Universidade de Aveiro
2010
Secção Autónoma de Ciências Sociais Jurídicas e Políticas

PAULO RICARDO LOPES BATISTA

DATA MINING NA IDENTIFICAÇÃO DE ATRIBUTOS VALORATIVOS DA HABITAÇÃO

Universidade de Aveiro
2010
Secção Autónoma de Ciências Sociais Jurídicas e Políticas

PAULO RICARDO LOPES BATISTA

DATA MINING NA IDENTIFICAÇÃO DE ATRIBUTOS VALORATIVOS DA HABITAÇÃO

Dissertação de mestrado apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Planeamento Regional e Urbano realizada sob a orientação científica da Doutora Gladys Castillo Jordan, Professora Auxiliar do Departamento Matemática da Universidade de Aveiro e do Mestre João José Lourenço Marques, Assistente da Secção Autónoma de Ciências Sociais Jurídicas e Políticas da Universidade de Aveiro

o júri

presidente

Doutor José Manuel Gaspar Martins
professor auxiliar da Secção Autónoma de Ciências Sociais Jurídicas e Políticas da Universidade de Aveiro

vogais

Doutor Carlos Manuel Milheiro de Oliveira Pinto Soares
professor auxiliar da Faculdade de Economia da Universidade do Porto

Doutora Gladys Castillo Jordan
professora auxiliar do Departamento de Matemática da Universidade de Aveiro

Mestre João José Lourenço Marques
assistente da Secção Autónoma de Ciências Sociais Jurídicas e Políticas da Universidade de Aveiro

agradecimentos

Aos meus orientadores, por todo o conhecimento e apoio que prestaram, tanto nas pequenas como nas grandes tarefas, determinantes para a concretização deste trabalho.
Aos meus pais, ao meu irmão e aos meus avós maternos pela força de todos os dias.
À Ana, pela companhia nos bons e maus momentos.
A todos os meus amigos e aos meus colegas de trabalho pela paciência no dia-a-dia.

Agradeço ainda à Janela Digital, pela disponibilização dos dados do portal Casa Sapo e à equipa do Laboratório Sapo da Universidade de Aveiro, que colaborou activamente na disponibilização e tratamento inicial dos dados do serviço Sapo Mapas.

palavras-chave

Data mining, habitação, mercado imobiliário, econometria, planeamento urbano

resumo

A teoria que sustenta o conceito de *preço hedónico* permitiu desenvolver uma ferramenta econométrica, simples e eficiente, para estudar o tema habitação a partir da informação associada às transacções no mercado. O desafio, associado à aplicação destes modelos, baseia-se na dificuldade de identificar, a partir dos volumes da informação actualmente existente, quais os atributos efectivamente determinantes para o processo de formação do valor de mercado.

A incapacidade de incorporar todos os atributos de uma habitação num modelo explicativo do preço não se deve, exclusivamente, às dificuldades e deficiências já conhecidas do funcionamento de mercado. As fontes de informação tradicionais têm disponibilidades de informação limitadas, reservando um papel chave ao conhecimento prévio do investigador. Este conhecimento é fundamental para o desafio de produzir nova informação ou de a recolher, a partir de dados existentes. Neste desafio, recentes técnicas de análise de dados, fornece ferramentas que complementam a recolha e selecção de atributos relevantes por parte de cada investigador.

A partir dos dois casos de estudo apresentados, pode concluir-se que a utilização de ferramentas de *data mining* permite reduzir, de forma mais eficiente que a utilização exclusiva de conhecimento empírico do investigador, o número de atributos necessários para explicitar a formação do preço da habitação.

Com a utilização destas ferramentas de análise de dados, a capacidade explicativa dos modelos, que identificam os determinantes do preço, não é afectada de forma substancial. Em muitos casos é possível melhorar a capacidade explicativa, pela eliminação de atributos que introduzem ruído e inconsistências no modelo econométrico. Noutros casos, demonstra-se que em problemas com maior complexidade, permite reduzir o número de atributos sem uma perda significativa de capacidade explicativa.

keywords

Data mining, housing, real estate market, econometry, urban planning.

abstract

The theory, behind the concept of hedonic price, enabled the development of a simple and efficient econometric tool. The challenge in applying hedonic models can be summarized in the identification of the attributes that are crucial to the market value.

The inability to incorporate all dwelling's attributes is not solely due to the difficulties and deficiencies associated with market operations. The traditional sources of information are limited. Prior knowledge plays a key role in the researcher's work. This is fundamental in the challenge of producing new information and collecting available data.

Recent techniques of data analysis provide new tools for collecting and selecting tasks.

From the case studies analysed, it is possible to conclude that data mining tools, with focus on the task of feature selection, allow to reduce the number of attributes needed to explain house prices.

The explanatory power of hedonic models is not substantially affected, by the use of a less number of attributes. In many cases, you can improve the explanatory power by eliminating attributes that introduce noise and inconsistencies in the econometric model. Concerning more complex problems, new selection algorithms allow reducing the number of required attributes, without a significant loss of the explanatory power.

Índice

Lista de Figuras

Lista de Tabelas

I. INTRODUÇÃO

I.1. – A relevância do tema habitação

A habitação é um tema incontornável para a esmagadora maioria da população, sendo considerado um dos bens básicos da sociedade. Suportado no recurso insubstituível solo[1] e, num enquadramento legal que garante o direito à propriedade privada, o tema habitação requer a consideração de múltiplos impactos – sociais, económicos e territoriais.

O estudo das problemáticas associadas a este tema surgiu com as transformações socioeconómicas associadas à Revolução Industrial nos séculos XVIII e XIX. Acompanhada de graves problemas de salubridade, colocou em causa a sobrevivência de parte significativa da população, originando uma preocupante deterioração da coesão social. A população, cada vez mais concentrada nas áreas urbanas, enfrenta problemas que despertam o interesse das ciências sociais. O desenvolvimento científico da economia, da sociologia e da geografia ocorreu, em grande medida muito associado ao estudo do tema habitação.

Paralelamente à importância do desenvolvimento científico nesta matéria, Correia (2002), refere a fundação da administração pública. Este sistema de organização da sociedade surge da necessidade da defesa do interesse social, o qual pressupõe a procura da ordem independentemente do poder político, a descriminação de regras claras não discricionárias e a necessária constância e coerência dessas regras, ao longo do tempo, que permitam o seu progressivo enraizamento e aperfeiçoamento, promovendo o desenvolvimento urbano mais harmonioso. É assim que à propriedade privada são colocados limites, abrindo espaço à intervenção de uma administração pública, pela

[1] O solo é efectivamente o suporte de qualquer edificação, tornando-se parte integrante desta. Naturalmente, não assume exclusivamente a função, por este facto, é comum a utilização da referência de mercado imobiliário, quando se refere uma transacção genérica de uma parcela de solo, como todo e qualquer bem que a esta esteja associado.
Neste trabalho, utiliza-se especificamente a designação mercado habitacional (ou da habitação) para o mercado de transacção de uma parcela de solo ocupada com a função habitação.

delegação de poderes dos cidadãos. O mecanismo tem como objectivo natural integrar todos os indivíduos no processo de desenvolvimento, garantindo regras e direitos sociais justos que promovam a coesão social.

A administração pública é hoje, o pilar do estado de direito. Destaca-se a sua actuação sobre as problemáticas da habitação: o quadro legal actual engloba múltiplas ferramentas legais que vão dos mecanismos de gestão e regulação (como a Lei dos Solos em Portugal, mas com equivalente em praticamente qualquer estado), os mecanismos de compensação e redistribuição (onde se destaca a tributação fiscal), e os instrumentos de planeamento e ordenamento territorial (onde em Portugal se destacam os planos municipais de ordenamento, com poder legal, responsáveis por definir as regras de ocupação do solo).

Apesar do poder da intervenção pública, a co-existência com o direito à propriedade, torna o solo um bem transaccionável, no mercado. As peculiaridades associadas ao funcionamento dos mecanismos deste mercado, que também se procura rever neste trabalho, resultam na prevalência de problemas como a descriminação social no acesso e usufruto condigno deste bem.

Em Portugal, com um desfasamento considerável em relação à maior parte do mundo industrializado, os problemas começaram a surgir mais tarde. Descontando as especificidades próprias do contexto nacional, o tipo de problemas e dificuldades enfrentadas não foram muito diferentes dos das restantes sociedades.

É no século XX que o tema habitação salta para o discurso dos actores políticos nacionais. Uma preocupação que se centrou durante todo o século no tradicional paradigma discursivo do «*direito à habitação»*. O actual texto constitucional (de 1974) continua a consagrar no seu artigo 65º, ponto 1, que «t*odos têm direito, para si e para a sua família, a uma habitação de dimensão adequada, em condições de higiene e conforto e que preserve a intimidade pessoal e a privacidade familiar*». Este preceito levou, nos últimos anos, à implementação de políticas públicas centradas nos apoios financeiros à aquisição de habitação, na afectação de novas áreas à ocupação urbanas e, inclusivamente, na produção pública de habitação.

Fortes transformações socioeconómicas, onde se destacam as novas estruturas familiares, aliadas à resolução global (ou parcial) das carências habitacionais mais gritantes a que se juntam outros factores exógenos, têm fundamentado a transformação

dos objectivos da intervenção pública. Do *direito à habitação* o discurso passou a centrar-se na necessidade de garantir o correcto equilíbrio entre oferta e procura, estabelecido pelo mercado. Guerra et al (2008), no Plano Estratégico da Habitação para o período de 2008 a 2013 (PEH), sublinha esta nova orientação. O plano refere que a necessidade quantitativa de alojamento é encarada como um problema pertencendo ao passado, sendo as actuais políticas habitacionais direccionadas para o desenvolvimento de respostas, a grupos sociais específicos. De forma mais geral, aponta que as preocupações devem dar atenção à regeneração urbana e, dependendo da gravidade das situações, nas intervenções para estabilizar os mercados habitacionais privados.

A rapidez das transformações sociais modernas e uma certa volatilidade, hoje atribuída ao mercado habitacional, traduzem-se num comportamento imprevisível dos agentes de mercado.

Neste sentido, vários investigadores têm alertado para a necessidade de desenvolver novos mecanismos de apoio à decisão. Carvalho (2003) refere que "*pensar a localização, a tipologia e o dimensionamento residencial é, agora como sempre, uma parte substancial do ordenamento, ganhando actualmente novas complexidades"*. O mesmo autor relembra que "*para que a política pública possa influenciar tipologias e sobretudo localizações é indispensável conhecer e partir das dinâmicas instaladas, reveladoras de lógicas e interesses de actores determinantes nessas transformações"*. No mesmo sentido, Correia (2002) alerta para o facto de que *os "insucessos do sistema de planeamento tradicional e a não observância das suas regras, resultam da dissociação entre a concepção dos espaços urbanos e a gestão dos planos*," preconizando que o "*processo de decisão só terá sucesso se"*, entre muitas actuações, "*procurar recolher e analisar cada vez mais informação sobre a realidade em que se pretende intervir"*.

É cada vez maior a capacidade de armazenar informação. Instituições públicas têm recolhido cada vez mais informação, estruturada de forma a responder às múltiplas funções que vem assumindo. No entanto, a estrutura de recolha e compilação de dados (como a desenvolvida pelo Instituto Nacional de Estatística - INE), começa a ser considerada limitada para as cada vez maiores necessidades de produzir informação – nomeadamente, pela necessidade de desenvolver informação cada vez mais desagregada territorialmente, que permita desenhar e avaliar políticas a uma escala local.

Contudo, existe um número cada vez maior de novas bases de dados, construídas por entidades privadas no âmbito das suas actividades comerciais. É previsível que estas possíveis fontes de informação não se encontrem estruturadas para a sua utilização aplicada aos problemas de investigação, facto que aponta a necessidade de implementar novas abordagens na análise de dados.

Os problemas que necessitam de resolução são vários e envolvem questões como:

- a acessibilidade à informação, condicionada pela sua dispersão pelos diferentes agentes, tornando-a fragmentada

- a dificuldade de cada actor envolvido utilizar de forma eficiente a informação de que dispõe, incluindo-se aqui as instituições públicas.

Um problema adicional apresenta-se ao investigador: perante maiores volumes de informação torna-se difícil a selecção do conjunto correcto de variáveis para um dado problema. Na generalidade, o conhecimento global do investigador é o ponto de partida para seleccionar um conjunto de variáveis a partir de todas as fontes disponíveis. No entanto, esse processo exige uma grande dependência da acumulação de conhecimento sobre o ao problema em análise, o que é um recurso limitado.

É normal que diferentes investigadores obtenham resultados diferentes se partirem de conjuntos de dados diferenciados. Dada a abrangência do tema habitação, assistimos à segmentação dos problemas de estudo. São vários os estudos que abordam temas específicos, limitando e simplificando os processos de recolha e selecção de dados.

Num cenário com cada vez maior volume de dados, estes desafios são cada vez mais importantes e uma tarefa aparentemente tão simples como identificar os atributos determinantes do preço da habitação envolve um complexo processo de selecção que o investigador só por si já não resolve de forma eficiente.

Centrado no tema habitação, o trabalho aqui apresentado terá como pano de fundo a demonstração das potencialidades da utilização de novas ferramentas para a selecção de atributos determinantes no preço da habitação. Espera-se que esta perspectiva permita demonstrar a capacidade de aprofundar ou alargar as áreas de investigação sem que ao investigador em temas territoriais seja requerida uma especialização nas ferramentas de análise de dados, em prejuízo da sua multidisciplinaridade científica.

I.2. – Objectivos e Metodologia

II.2.1. Objectivos

Este trabalho pretende constituir-se como um contributo para o processo de análise das dinâmicas do mercado imobiliário em Portugal.

A relevância do mercado, numa altura em que os agentes públicos optam, cada vez mais, por delegar nos seus mecanismos eficientes a supressão das necessidades de habitação, torna importante a existência de ferramentas que permitam perceber o processo de formação de preços. Como veremos, sendo a habitação um bem tão importante mas também tão complexo, identificar os atributos determinantes do preço da habitação é uma contribuição simples mas que envolve importantes desafios.

Existindo uma diversificada e consistente literatura científica que fornece ferramentas para a concretização do objectivo genérico atrás enunciado, procura-se construir um modelo metodológico conceptual para identificação dos principais atributos determinantes do preço da habitação em Portugal, tirando partido de fontes de informação existentes. Procura-se explorar os dados públicos, disponibilizados pelo Instituto Nacional de Estatística (INE) e ir mais longe com a utilização de bases de dados, disponíveis em portais como o Casa Sapo e o Sapo Mapas.

A dissertação procura, desta forma:

i) caracterizar o fenómeno de formação do preço da habitação e assinalar as deficiências e dificuldades no tratamento da informação existente para este fim;

ii) demonstrar as potencialidades de integrar novas fontes de informação, desenvolvidas em contextos diferenciados, no sentido da melhoria e eficiência das análises do fenómeno de formação de preços;

iii) explorar e testar diferentes abordagens para a identificação dos determinantes de formação do preço no mercado habitacional;

Estes objectivos culminam na tentativa de identificar e caracterizar os factores determinantes do preço da habitação em Portugal, para diversas escalas de análise.

Na concretização destes objectivos gerais procura-se de forma complementar:

i) desenvolver uma metodologia que permita apoiar o investigador na selecção do número restrito de atributos;

ii) identificar a importância hierárquica de atributos determinantes do preço da habitação;

iii) testar a versatilidade de novas abordagens, nos processos de análise de dados, com objectivos de criar ou sintetizar.

II.2.2. Metodologia

Para atingir os objectivos propostos propõe-se os seguintes passos para a elaboração do percurso metodológico:

- **Definição e delimitação do objecto de estudo:** enquadrado no conjunto de objectivos, descrever e delimitar a problemática permitirá identificar os desafios ao estudo do mercado imobiliário. Facultará a possibilidade de descrever e identificar as exigências de diferentes modelos econométricos actualmente propostos pelos investigadores.
- **Recolha de informação:** recolher e catalogar a informação disponível, envolve um processo de recolha e (re)estruturação dos dados que permita a aplicação das técnicas exigidas pelas ferramentas que nos permitem a construção dos modelos seleccionados.
- **Selecção de ferramentas:** a aplicação dos modelos econométricos requer um trabalho de identificação e selecção das ferramentas que os permitem concretizar. Os critérios baseiam-se, para além dos elementos teóricos, na relação simplicidade versus capacidade explicativa que cada ferramenta proporciona. Esta etapa engloba o necessário trabalho de apreender e desenvolver as ferramentas necessárias para a aplicação dos modelos seleccionados (aquisição de novos conhecimentos, definição de requisitos de software, entre outros).

- **Tratamento e modelação:** a construção de modelos associados a um dado fenómeno em estudo implicam lidar com a necessária complexidade da informação existente e com a capacidade de extrair e limitar os conceitos que se pretendem estudar. É expectável a necessidade de estabelecer um conjunto de casos de estudo que permita conciliar a complexidade dos modelos, a capacidade de os aplicar e a concretização dos objectivos.
- **Análise e conclusão:** os resultados obtidos permitirão avaliar a concretização dos objectivos estabelecidos.

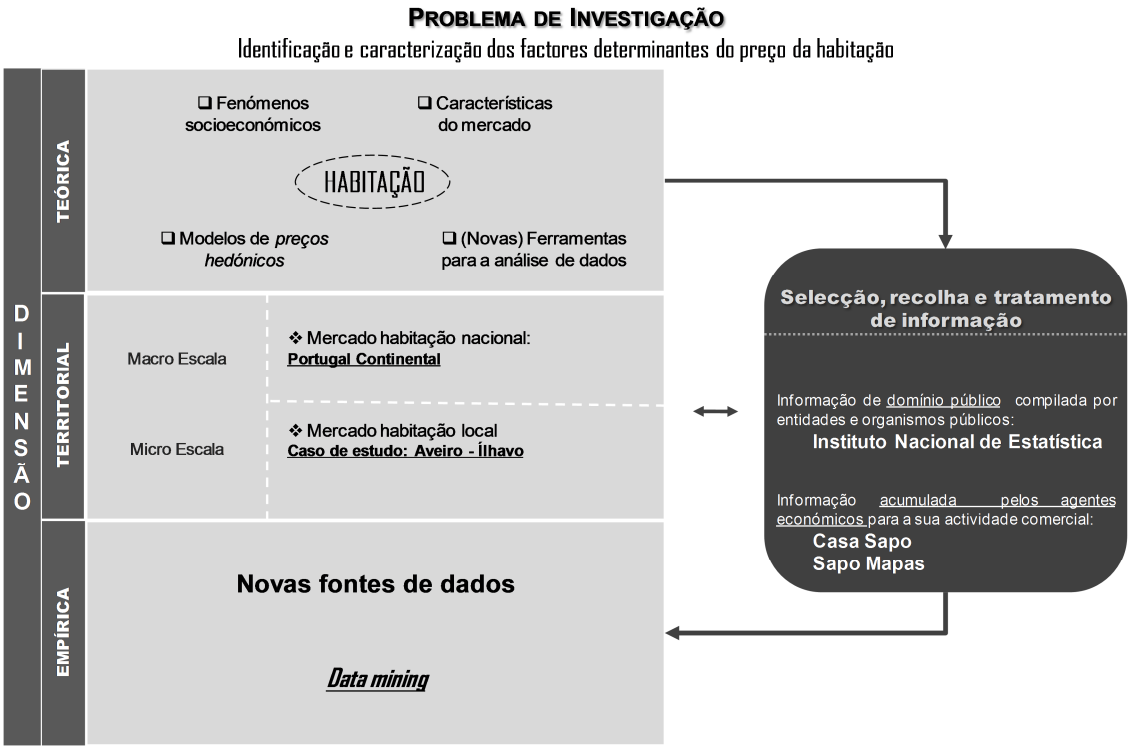

Figura 1 Interacção entre a metodologia proposta e as dimensões de análise que se pretendem desenvolver

A metodologia proposta procurará expor resultados em 3 dimensões de análise (Figura 1):

Dimensão teórica: a consulta de elementos bibliográficos terá como objectivo balizar a problemática associada ao estudo da habitação, nomeadamente: descrever as características do bem habitação que determinam a formação do preço de mercado; identificar, as formas de quantificar essas mesmas características; analisar os modelos econométricos que permitam responder, de forma eficaz, à quantificação dos seus impactos no preço da habitação.

Dimensão territorial: a habitação tem uma óbvia dimensão territorial intrínseca. As diferentes delimitações espaciais do problema implicam considerar os problemas e desafios associados a cada uma destas escalas de análise.

Dimensão empírica: ao nível empírico procuram-se identificar as vantagens e desvantagens dos modelos e ferramentas escolhidos para tratamento da informação disponível e o seu ajustamento aos objectivos propostos. Com a expectável complexidade associada aos volumes de informação disponíveis, o trabalho centrar-se-á na abordagem de casos de estudo exemplificativos.

II. REFLEXÃO TEÓRICA

Neste capítulo resumem-se as mais importantes observações teóricas que sustentam a construção deste trabalho e dos respectivos estudos empíricos. Tem como objectivo enquadrar a selecção das ferramentas econométricas utilizadas.

O capítulo inicia-se com a identificação dos principais fenómenos socioeconómicos com impacto directo nos padrões sociais e territoriais que caracterizam a temática da habitação em Portugal.

A segunda secção complementa a abordagem socioeconómica com a visão, centrada na economia, das características da habitação que a destacam de outro tipo de bens transaccionáveis. Procura-se descrever a forma como estes aspectos económicos são responsáveis por induzir determinados comportamento e, consequentemente, influenciar os mecanismos de formação do preço de mercado.

Apesar da dificuldade em estudar o mercado da habitação, a sua relevância social tem sido responsável pela atenção dada pela teoria económica ao desenvolvimento de modelos para o estudo do mercado imobiliário e da habitação em particular. São as teorias económicas e a observação de aspectos do funcionamento dos mercados que marcam a construção das ferramentas hoje utilizadas para estudar o tema habitação. A terceira secção enquadrará as teorias económicas que sustentam os novos modelos e ferramentas.

Por fim, destaca-se a selecção, para este trabalho, dos *modelos de preços hedónicos*. Sendo um exemplo do tipo de ferramentas que procuram tirar partido de informação criada no momento da transacção, permitem identificar a relevância de diferentes atributos na formação do preço da habitação – o que é o objectivo principal deste trabalho.

II.1 – Fenómenos socioeconómicos e habitação

Várias disciplinas científicas debruçam-se sobre o tema habitação. Destacam-se os trabalhos que abordam a habitação numa perspectiva alargada, inserindo as suas questões no âmbito das dinâmicas da sociedade. Nesta secção, procura-se apresentar os mais importantes fenómenos socioeconómicos das últimas décadas. Este breve enquadramento permite identificar as questões chave da habitação à escala macro, de Portugal continental. São estas características e grandes padrões territoriais que nos indicam quais os expectáveis atributos que caracterizam a habitação em Portugal.

Movimentos Migratórios

As migrações são fenómenos sociais muito importantes, ocorrendo em intervalos de tempo curtos quando comparados com a durabilidade da habitação. Induzem mudanças nas características da habitação ao alterarem significativamente a estrutura social e, consequentemente, as necessidades habitacionais. Em Portugal, Valença (2001) descreve dois acontecimentos chave:

> Migração dos anos 50 e 60.
>
> A emigração para o exterior é, neste período, responsável pelo movimento de aproximadamente 10% da população residente. O fenómeno foi intenso nas áreas rurais, onde as condições de vida eram comparativamente piores.
>
> Internamente, em consequência das tímidas tentativas de industrialização económica, assistiu-se também à deslocação de população das áreas rurais para as áreas urbanas.
>
> Ambos os processos migratórios acentuaram os desequilíbrios pré existentes entre áreas rurais e urbanas e, essencialmente, entre o interior e o litoral.
>
> Período revolucionário de 74.
>
> Com a transformação política associado ao 25 Abril de 1974, surge, num curto espaço de tempo de 2 a 3 anos, um afluxo de portugueses, até aí residentes nas ex-colónias. Segundo Valença (2001), a magnitude deste fluxo compensa quantitativamente o movimento de saída descrito no ponto

anterior. Contudo, mais uma vez, a fixação dos novos residentes ocorre nas áreas de maior dinamismo económico – localizadas no litoral e de onde se destacavam as áreas metropolitanas de Lisboa e Porto.

Como resultado, agravam-se ainda mais os problemas de desequilíbrio territorial rural / urbano e interior / litoral.

Estes dois acontecimentos são factores explicativos de um espaço rural maioritariamente abandonado. Com o consequente abandono e degradação de uma parte significativa do *stock* de habitação. Em contrapartida, associado ao curto espaço de tempo e ao volume assinalável do fenómeno migratório, verificou-se um grave problema de escassez, muito significativo nas áreas urbanas, e em especial, nos dois pólos metropolitanos: Lisboa e Porto.

Desenvolvimento económico

O fraco desenvolvimento económico e as precárias condições financeiras do estado ao longo dos anos, traduzem a persistência de um pronunciado diferencial de qualidade de vida que apenas nas últimas décadas se tem atenuado. Este diferencial não é exclusivamente territorial (Portugal / exterior ou rural / urbano ou interior / litoral), sendo, também, um fenómeno de desigualdade social.

Os fenómenos migratórios das décadas de 50 e 60 permitiram uma (débil) dinamização económica, justificável com a entrada de divisas, enviadas pela comunidade emigrante, e pela concentração de mão-de-obra (barata) nas áreas urbanas. O desenvolvimento ocorre nos territórios mais dinâmicos, servidos por melhores vias de comunicação, com maior dimensão e consequentes ganhos em escala e com uma estrutura económica apta a abarcar novos investimentos. Nas áreas rurais, dominava e domina um sector primário, tradicional.

As consequências de um desenvolvimento socioeconómico diferenciado ao nível territorial sentem-se, ainda hoje, em várias áreas. No domínio da habitação, Valença (2001) refere que, a partir dos anos 80, o grave desequilíbrio entre as necessidades de habitação (qualitativamente e quantitativamente elevadas nas áreas urbanas) e o elevado *stock* existente (de características qualitativamente baixas, localizado nas áreas rurais, cada vez mais desertificadas) emergiu.

Intervenção da administração pública

Numa outra vertente, a análise da intervenção da administração pública nas respostas aos desequilíbrios provocados pelos fenómenos socioeconómicos, permite concluir que esta foi incapaz de inverter ou estancar os múltiplos problemas que brotaram. É comum que as justificações apontem para a falta de capacidade financeira e a inadequação do sistema político. Tal como refere Valença (2001), num país que se caracteriza por um sistema de provisão habitacional quase totalmente na mão da iniciativa privada, esta deu-se a reboque do desenvolvimento económico, em especial no que se refere à localização e às formas de ocupação.

A ausência de uma intervenção pública eficaz traduziu-se ainda no aparecimento do fenómeno de construção de habitação de génese ilegal. Formaram-se bolsas de habitação precária e expandiram-se áreas de construção ilegal ocupadas por classes mais abastadas, mas com dificuldade em aceder ao mercado. Como veremos mais à frente, os mecanismos de mercado da habitação não são propriamente eficientes por natureza, justificando a sua incapacidade de se adaptar às reais necessidades da população.

A integração, na então designada Comunidade Económica Europeia (CEE), em 1986, revelou-se um factor atenuante. A política de desenvolvimento regional das regiões menos desenvolvidas da comunidade, tornou-se a fonte do dinamismo económico e proporcionou alguma capacidade financeira ao estado. Os esforços da política pública direccionaram-se para a resolução dos problemas mais gritantes, investindo em programas de realojamento da população residente em habitações precárias. Ao mesmo tempo, procurou-se intervir no mercado através da disponibilização de terrenos públicos para projectos cooperativos, de habitação social e a custos controlados, bem como na expansão das áreas urbanizáveis.

A oferta de solo com possibilidade construtiva cresceu de forma exponencial e, a meados da presente década permitiria, caso todas as áreas urbanas e urbanizáveis fossem efectivamente ocupadas, albergar um valor várias vezes superior ao da população residente. De forma complementar assistiu-se a um forte investimento público na melhoria generalizada das infraestruturas: dos serviços básicos como água, luz e saneamento a outras, com impactos muito directos no dinamismo do mercado imobiliário, como é o caso das infraestruturas de transportes.

Contudo, como refere o recente *diagnóstico de carências habitacionais* desenvolvido no âmbito do Plano Estratégico da Habitação - PEH (Guerra et al, 2008), estas intervenções não foram suficientes para resolver definitivamente os problemas existentes.

Os padrões actuais da habitação

As últimas décadas ficaram marcadas pela forte expansão do mercado da habitação. Por exemplo entre 1970 e 2001 registou-se uma duplicação do número de habitações existentes, sendo o ritmo de crescimento do *stock* habitacional sempre superior ao ritmo de crescimento do número de famílias.

Em 2001, para os 10 milhões de habitantes, organizados em cerca de 4 milhões de famílias, existiam cerca de 5 milhões de habitações. Destas, 73% eram ocupadas como residências habituais e as restantes 28% como segundas residências ou encontravam-se desocupadas. No entanto, os problemas estavam longe de estar resolvidos e os indicadores disponíveis apontam para a prevalência de um conjunto de problemas. O PEH dá destaque a dois grandes problemas:

i) *a degradação dos alojamento*: com 40% dos alojamentos a necessitarem de algum tipo de reparação, fica demonstrado que, por alguma razão, grande parte do *stock* de habitação encontra-se abandonado.

ii) *a sobrelotação*: atinge 16% do total de alojamentos e alerta para persistentes dificuldades no acesso à habitação

Como explicações para estes dois fenómenos são apontadas duas novas razões, que complementam as consequências dos três fenómenos socioeconómicos históricos atrás descritos: o persistente baixo rendimento da maioria das famílias e a ausência de uma oferta de habitação ajustada às efectivas necessidades da procura.

Os números de pobreza são muito elevados. Encontramos mais de 20% da população portuguesa abaixo do limiar de pobreza. Segundo o PEH, a acessibilidade ao alojamento obtida através da propriedade representa um maior impacto no rendimento das famílias. O encargo mensal para estas famílias é avaliado em 32% do rendimento disponível, sendo de 66% nas famílias mais pobres. Ao contrário do que seria de

esperar, a propriedade é a forma dominante de acesso à habitação, com valores em torno do 75%, inclusive na população mais pobre.

Por outro lado, também o modelo baseado na construção e venda de habitação dá sinais de ruptura. Os discursos começam a mudar e associações profissionais ligadas ao ramo imobiliário apresentam alguns estudos que, como por exemplo o relatório sobre o mercado da reabilitação (Martins, 2009) editado pela Associação de Empresas de Construção, Obras Públicas e Serviços, aponta para uma nova direcção estratégica: o investimento em reabilitação como o futuro do sector. As estimativas apresentadas avaliam as necessidades globais de reabilitação em torno dos 200 mil milhões de euros.

A resposta a estes fenómenos e a resolução destes desequilíbrios passará por uma maior e mais eficiente capacidade de intervenção pública. A informação sobre o mercado da habitação é um aspecto crucial.

Identificar os determinantes no processo de formação do preço da habitação é o primeiro passo para uma correcta monitorização dos mercados imobiliários e para o desenvolvimento de estudos, por parte da administração pública.

Estes simples indicadores têm impactos múltiplos. Podemos destacar a sua utilidade em mecanismos como a determinação do IMI[2] (Imposto Municipal sobre Imóveis), permite, por si só, tornar este imposto mais justo e mais eficaz nos objectivos sociais com que foi implementado. Como outros exemplos destaca-se o contributo para as ferramentas de avaliação custo benefício: o projecto de investigação Custos e Benefícios de uma Ocupação Dispersa[3] estuda o custo associado a uma ocupação dispersa que é comum ser um fenómeno classificado como nefasto para os interesses globais da comunidade. Contudo, esta classificação à priori esquece frequentemente os benefícios associados a esta ocupação e que explicam a sua expansão: o facto de nestes locais existir o tipo de habitações que as pessoas efectivamente procuram. Assim, conhecer quais os atributos que as pessoas valorizam serão um aspecto fulcral, para desenhar soluções urbanas mais compactas e que permitam um equilíbrio maior entre os custos e os benefícios proporcionados à população.

[2] O IMI é considerado uma fonte de receitas muito importante para o financiamento da administração local. Sendo ajustado ao valor patrimonial estimado para cada imóvel, pode ser considerado um imposto progressivo.
[3] Mais informações em http://www.ua.pt/ii/ocupacao_dispersa/

II.2 – Características do mercado da habitação

O primeiro aspecto a ter em conta no estudo do mercado habitacional é relativo às **características intrínsecas da habitação** que a diferencia de outros bens transaccionáveis.

Desde logo, a habitação assenta na utilização intensiva do recurso solo sendo os seus atributos indissociáveis deste. Destacam-se como principais aspectos genéricos desta relação:

- a heterogeneidade: dada pela existência de atributos de natureza irrepetível para cada parcela de solo, tornando o bem habitação um sucedâneo imperfeito; por exemplo, os atributos de localização, dimensão, forma, características geotécnicas, características paisagísticas e de vizinhança de cada parcela de solo são únicas.
- a imobilidade: traduz a incapacidade de deslocar o bem e usufruir dos seus serviços ou funções em localizações diferentes; este aspecto tem consequências ao nível da procura e da oferta: estas dinâmicas tornam-se referenciadas a uma determinada localização, provocando a possível criação de mercados imobiliários localizados (e diferenciados). Como consequência, aumentam as dificuldades dos diferentes agentes percepcionarem o funcionamento do mercado global e ainda mais dos possíveis submercados que comporta.
- a durabilidade**:** a duração de uma habitação é potencialmente elevada, dependendo de múltiplos factores e não se esgotando após o uso; acresce que a durabilidade do solo tende para infinito. Estes aspectos significam:

  - um reduzido fluxo de stock renovado perante o stock acumulado, dificultando ajustes às dinâmicas de procura;

  - um comportamento pouco previsível dos agentes – uma qualquer decisão sobre um bem durável é mais facilmente adiável, visto que o bem conserva as suas características ao longo do tempo, sendo expectável a variação, acentuada e cíclica, nos volumes de ambos os stocks.

As características da habitação atrás descritas expõem parte da complexidade associada ao funcionamento do mercado imobiliário. Acresce que são características que influenciam de forma determinante a **motivação para a participação no mercado.** Para lá das tradicionais motivações económicas, existem três aspectos com uma consequência importante no mecanismo de formação de preços:

i) Habitação como bem de consumo e simultaneamente de investimento. A genérica durabilidade da habitação permite suportar dois tipos de procura que se encontram muitas vezes separadas: o usufruto e o investimento de capitais. Neste último aspecto a expectativa de rentabilidade por parte dos agentes de mercado, na comparação com outras formas de investimento, é frequentemente superior.

ii) Forte valoração. A natureza única dos atributos associados a cada parcela de solo e a cada habitação em si é um dos factores que explicam os preços altos relativos à real capacidade de investimento dos agentes de mercado. Uma consequência importante desta atitude é a sensibilidade das condições de mercado às oscilações das dinâmicas económicas gerais.

iii) Associação de status. As características de cada parcela permitem a afirmação de riqueza e de capacidade económica do seu proprietário. Uma consequência óbvia é que a manutenção deste *status* é responsável pela retenção de grande parte do *stock* existente.

Neste quadro, as **operações de mercado** são caracterizadas por elementos importantes. O conceito de propriedade, que segundo vários autores decorre da apropriação pelo ser humano das coisas que foi encontrando e de que foi dispondo, com vista à satisfação das suas necessidades, é reconhecida e valorizada na esmagadora maioria dos sistemas de organização social. Assim reconhece-se um conjunto de restrições naturais:

- Ocorrência de transacções na presença de parcelas desocupadas**.** Responsável principal pelo elemento especulativo no processo de formação de preços: a expectativa de valoração, ou a reacção perante a desvalorização, é um critério central da acção dos proprietários de parcelas desocupadas. Os proprietários são ainda propensos a adiar o momento em que entram no mercado devido à durabilidade associada à parcela de solo e da habitação (tal como já referido

em ponto anterior). Este aspecto contribui para a dificuldade em quantificar situações de escassez ou de excesso na oferta: a facilidade com que as parcelas podem entrar e sair do mercado [em função da referida atitude especulativa, por exemplo] provoca uma alteração rápida no seu equilíbrio, contribuindo para maior volatilidade.

- Usufruto da propriedade sujeito a requisitos legais. Consequência das funções atribuídas à administração pública, que inclui poderes:

  - *normativos*: por exemplo, os planos de ordenamento, que estabelecem usos permitidos e capacidades construtivas a cada parcela.

  - *de investimento*: por exemplo, sendo responsável pela execução de infraestruturas básicas.

  - *de licenciamento*: com impacto nos custos de transacção e que procuram a prossecução do *bem comum*., garantindo que o usufruto da propriedade

  Os requisitos legais acentuam os aspectos da heterogeneidade de cada.

- Existência de externalidades. A acção da administração pública, por exemplo, é naturalmente uma das principais fontes de externalidades (embora não a única): o acto normativo, de investimento ou de licenciamento atribui benefícios e penalizações que diferenciam as parcelas. Neste quadro surge a capacidade de influência que determinados agentes podem adquirir no processo de decisão da administração pública e que pode resultar em benefícios próprios inapropriados. No entanto, as externalidades são associadas a outros factores mais comuns: o tipo de ocupação e de usufruto legalmente concedido de uma parcela de solo causa impactos nas parcelas vizinhas; a título de exemplo quando uma determinada parcela de solo é licenciada para uma actividade com grandes níveis de poluição, esse facto pode influenciar negativamente a valoração das parcelas limítrofes, especialmente as ligadas à utilização habitacional.

- Intervenções esporádicas. Os agentes económicos intervêm no mercado esporadicamente., o que se deve, por um lado a uma necessidade naturalmente fortuita de habitação e, por outro, à dificuldade de percepção das complexas dinâmicas do mercado. Acresce que a intervenção dos investidores e promotores não é só devida à inerente dinâmica da procura

> como à tentativa de captação de externalidades que permitam maximizar o rendimento expectável. A intervenção reduzida é comummente apontada como uma das principais causas das inúmeras assimetrias de informação existentes entre os diferentes agentes, tornando as transacções pouco transparentes.

Do quadro de limitações ao funcionamento do mercado ressalta uma vincada natureza de funcionamento imperfeito. A administração pública, no seguimento dos seus objectivos de protecção do bem comum, procura intervir de forma eficiente na correcção destas imperfeições, devendo apontar, segundo Carvalho (2003), para a adopção de acções que permitam tornar o mercado mais transparente. Estas acções envolvem necessariamente a necessidade de organizar e divulgar informação sobre as transacções e respectivos preços – um facto que também aqui pretendemos desenvolver. Para este objectivo, a teoria económica tem proporcionado o desenvolvimento de vários modelos de análise que, como veremos nas próximas secções, são razoavelmente simples nas suas concepções, ao mesmo tempo bastante eficazes para problemas que se prendam com a obtenção de informação.

II.3 – Teoria económica na análise do mercado imobiliário

Um dos primeiros problemas de investigação da ciência económica centrou-se na análise do fenómeno de formação da renda fundiária[4]. Adam Smith, considerado um dos pioneiros da ciência económica, elaborou a primeira teoria que permitia explicitar os valores diferenciados do solo. Apoiado no estudo da produção agrícola, atribuía a formação de renda fundiária diferenciada aos custos de produção de um mesmo produto agrícola quando realizado em diferentes parcelas com características de solo (o factor produtivo) distintas, deduzida dos custos de mão-de-obra.

O trabalho de Adam Smith serviu como o primeiro argumento em favor do controlo social sobre a apropriação individual das mais-valias e das limitações dos direitos decorrentes da propriedade do solo. Iniciou a discussão sobre a necessária intervenção de entidades pública como forma de garantir equidade e justiça social.

Numa perspectiva complementar surgem os contributos da teoria da economia espacial. Albergaria et al (2010) destacam os trabalhos de Von Thünen, Christäller e Lösh que se centram no factor distância ao(s) mercado(s) como explicativos dos padrões de distribuição espacial da produção. Estes modelos vêm complementar a teoria de Adam Smith, oferecendo uma explicação mais sólida do fenómeno de formação de renda fundiária ao nível macro. As limitações surgiram quando o interesse da investigação se começou a centrar em territórios mais pequenos, ao nível do perímetro urbano, por exemplo, onde a delimitação do mercado é mais ambígua.

O desenvolvimento da teoria económica trouxe novos conceitos com impactos importantes para os estudos da formação do valor do solo. O conceito utilidade é um dos mais relevantes. Permitiu assumir que, tal como outros bens, o valor do solo pode ser determinado como o resultado de uma função que traduz o conjunto de bens e serviços que a sua utilização proporciona. Esta definição mais abrangente é, contudo, mais complexa. Os bens e serviços proporcionados por uma dada parcela têm,

[4] O significado de renda aqui utilizado refere-se à *quantia recebida como benefício* pelo usufruto de um dado bem.

necessariamente, configurações diferentes para cada um dos agentes económicos que intervém no mercado, lançando novos desafios de investigação.

Alfred Marshall é apontado como um dos primeiros autores que menciona a concepção de utilidade de um bem num modelo teórico explicativo do processo de formação da renda fundiária. Partindo dos modelos anteriores, Marshall assume que a utilidade do solo é determinada tanto pela quantificação do valor da localização como do valor intrínseco de uma parcela de solo. Argumenta que por este motivo a renda fundiária tanto poder ser maior quanto mais imperfeito for o mercado, isto é, quando para certas parcelas não existem alternativas comparáveis em termos de satisfação da procura sobressai o fenómeno de raridade, de limitação da oferta e de satisfação da procura que deve ser introduzido nos modelos explicativos do valor do solo.

A reflexão de Marshall sugeria ainda a impossível distinção entre o mercado de solos e o mercado de habitações. Ambos são parte integrante do valor de uma dada parcela, não podendo ser dissociados nos seus pressupostos visto que partilham os mesmos mecanismos de mercado.

Na senda do conceito de utilidade, Correia (2002) refere os trabalhos de Hurd, Haig e Ely que numa perspectiva puramente descritiva dos processos de formação do preço de mercado, identificam inúmeras imperfeições. Estes trabalhos permitiram ressaltar aspectos exógenos, identificando relações de dependência no comportamento de diferentes agentes económicos perante as vantagens relativas de diferentes localizações, referem exemplos como o dinamismo demográfico, o desenvolvimento das redes de transporte, o desenvolvimento de serviços públicos, entre outros.

Na linha de investigação da identificação de aspectos exógenos ao mercado imobiliário surgem os estudos referentes a aspectos psicossociais. Correia (2002) realça, neste aspecto, os trabalhos de Halbwachs. Inserindo a sua investigação no que designou por "factor situação" (de uma rua, de um bairro), propôs o valor de opinião como resultado de influências sociais que não têm necessariamente a ver com necessidades reais – ou seja, a influência das características socioeconómicas da vizinhança. Por exemplo refere, a influência positiva ou negativa no valor do solo da concentração de grupos / classes sociais ou determinadas actividades económicas.

Por volta dos anos 50 do século XX, Turvey e Ratcliff apresentam modelos para o processo de formação do preço do solo com a inclusão de diversas características

identificadas ao funcionamento do mercado. A dificuldade em obter e relacionar os dados não permitiu realizar demonstrações quantitativas, mas apontavam como pistas a utilização de indicadores como o rendimento e estudos sobre as expectativas de valorização futura de um investimento como as principais restrições a serem incluídas no modelo, para além dos já comuns e aceites factores físicos e de localização.

Os modelos de Turvey e Ratcliff introduziam contudo uma nova teoria. Apontavam que o resultado da necessária concretização deste tipo de modelos seria a verificação de uma expectável segmentação do mercado.

Correia (2002) refere Wendt como o responsável por propor a utilização dos preços de mercado. Propondo um modelo quantitativo em que procura integrar os principais factores do valor do solo, enfrentou dificuldades em quantificar as variáveis que pretendia incluir no modelo, obrigando-o a utilizar valores teóricos. Não sendo capaz de testar a sua ferramenta com dados reais, deu o tiro de partida para o desenvolvimento das ferramentas que permitiam quantificar os modelos propostos. O ressurgimento da microeconomia tornaram estas tentativas ainda mais relevantes, levando investigadores como Wingo e Alonso, na década de 60, a proporem novos modelos quantitativos que, apesar de incompletos, forneceram as primeiras ferramentas válidas para um modelo geral do valor do solo.

O insucesso relativo das abordagens quantitativas até então desenvolvidas é justificado com a necessidade de simplificação dos modelos propostos, de forma a contornar os constrangimentos técnicos das ferramentas de cálculo e da informação disponível.

Um novo contributo surge com a *Teoria do Consumo de Lancaster* (1966). Lancaster defende que num mercado em concorrência perfeita, no qual o processo de formação de preços é transparente, o valor de um bem é a agregação do valor dos seus atributos. Assim, por exemplo o valor de uma camisola pode ser calculado pela agregação do valor de atributos como a cor, a natureza do material, a necessidade do comprador, entre outros atributos que, se a transacção ocorrer em concorrência perfeita, estão identificados.

A simplicidade da teoria proposta por *Lancaster* despoletou o desenvolvimento de um grande número de ferramentas. Segundo Marques (2010), é possível identificar dois tipos de abordagens comuns nos trabalhos de investigação desenvolvidos desde então.

Uma abordagem designada por *preferências declaradas*, procura simular o processo de escolhas no mercado através de métodos como o inquérito. As preferências individuais por um dado bem podem ser determinadas através da avaliação da vontade para pagar por um conjunto de hipotéticas habitações, às quais está associada uma variação implícita de atributos seleccionados pelos investigadores. As preferências são determinadas pelas diferenças quantitativas estabelecidas pelos inquiridos. Uma alternativa a este tipo de aplicação é o questionário onde o inquirido avalia isoladamente cada um dos atributos.

A cada vez maior acessibilidade a informação associada a uma transacção proporcionou o desenvolvimento de ferramentas designadas por "*preferências reveladas"*. Tendo como grande vantagem não necessitarem de produzir os dados que permitem estudar um determinado problema, tiram partido directo da *teoria do consumo de Lancaster*. A informação de mercado permite determinar e identificar os diversos atributos associados a um dado bem transaccionado. Das ferramentas desenvolvidas destacam-se hoje as associadas à construção de *modelos de preços hedónicos*, que se procurará explorar de forma mais pormenorizada.

II.4 – Modelos de preços hedónicos

II.4.1. Conceptualização

Os *modelos de preços hedónicos* foram desenvolvidos de forma independente da teoria económica, sendo comum apontar os trabalhos de Court, em 1939, para a indústria automóvel, como os principais propulsores desta técnica. Com o objectivo de construir um índice de preços, a técnica aplicada por Court tirava partido da disponibilidade de grandes volumes de informação associada a cada modelo de veículo produzido. Cruzando os atributos de cada veículo com os respectivos preços de venda, o índice determinava o impacto nas vendas da introdução de atributos específicos nos automóveis. A técnica permitia ainda quantificar o valor desses atributos constituindo uma ferramenta importante na estratégia comercial das construtoras automóveis. A técnica baseia-se na existência de uma variedade de preços associada à diversidade de características de um determinado bem.

Desenvolvida num contexto empresarial específico existiam algumas limitações que dificultaram a sua generalização:

i) Ausência de uma fundamentação teórica consistente no âmbito da econometria.

ii) Limitações de recolha de informação e dificuldades de efectuar os cálculos necessários.

A introdução da *teoria do consumo de Lancaster* e a maior disponibilidade de informação e de meios técnicos permitiram o pioneiro trabalho de Rosen, em 1974. Apontado como o primeiro *modelo de preços hedónicos* para determinação do valor de atributos intrínsecos da habitação, Rosen justifica a sua metodologia invocando a concepção de Lancaster, onde uma habitação pode ser analisada a partir do conjunto de atributos que a compõe. Assim, o valor económico dos atributos é determinado pela relação entre os preços observados e o conjunto de atributos que compõe cada habitação.

O valor de mercado da habitação dependerá das preferências dos consumidores pela combinação de atributos associada a cada habitação e pela concretização dos

objectivos de maximização do lucro, dos proprietários vendedores. Para construir este modelo, impõe-se a utilização de dados de transacção, num mercado em equilíbrio – que permite associar as diferenças de valor de cada habitação ao valor de cada um dos atributos.

Os dados necessários para construir um *modelo de preços hedónicos* da habitação são cada vez mais fáceis de obter. Se os estudos iniciais estavam limitados pela disponibilidade de informação estatística agregada, as bases de dados actuais armazenam volumes de informação altamente detalhados. Esta informação torna estes modelos ainda mais eficientes, possibilitando identificar e testar um cada vez maior número de transacções e, associado a cada habitação, um maior número de atributos.

II.4.2. Formulação do modelo hedónico

A decisão do consumidor baseia-se na optimização da escolha de um conjunto de atributos espelhados numa habitação que satisfaz as suas necessidades. Este processo de optimização exige uma quantidade significativa de bens no mercado, substitutos perfeitos.

O valor global de uma habitação num mercado em equilíbrio pode então ser representado como uma função:

$$p\ (\mathbf{X}) \qquad \textbf{[1]}$$

sendo:
p – preço da habitação
$\mathbf{X} = (X_1, X_2, \dots, X_n)$ – habitação $\mathbf{X}$ caracterizada pelos atributos $X_1, X_2, \dots, X_n$

Cada consumidor difere na avaliação de um mesmo bem, estabelecendo uma relação subjectiva de dependência própria entre os atributos de uma habitação e o seu valor global. O mecanismo de equilíbrio entre oferta e procura traduz a determinação do valor de transacção da habitação.

Assim, conhecidos os dados de cada transacção é possível construir um modelo matemático que traduza a relação dada em [1] e, consequentemente, determinar o valor dos diferentes atributos das habitações no mercado.

O primeiro desafio associado à construção de modelos de *preços hedónicos* prende-se com a escolha do tipo de função matemática que se deverá considerar para determinar a relação preço - atributos. Malpezzi (2008), em linha com diversos autores, descreve as diversas especificações que têm sido testadas para a formulação matemática que passam pela clássica regressão linear, pela utilização de formulações lineares alternativas como a log-linear, semi-log, ou ainda novos tipos de formulações como a transformação de Box-Cox[5], etc.

Não existe uma imposição da teoria económica relativa à formulação matemática óptima. Utilização de critérios comparativos, baseados em medidas de qualidade do ajuste do modelo e na análise da significância estatística dos coeficientes estimados assumem-se frequentemente como bons critérios de escolha. O teste t, para mensuração dos níveis de significância dos coeficientes das variáveis independentes e o coeficiente de determinação R^2 são indicadores comparativos amplamente utilizados no contexto das ciências sociais e que também se podem assumir como bons critérios.

Matematicamente, a formulação mais simples para este problema consiste na definição de um modelo de regressão linear. O *modelo de preços hedónicos* pode ser expressado como:

$$p\,(\mathbf{X}) = \mathbf{b}\,\mathbf{X} + \varepsilon \qquad [2]$$

sendo:

$p\,(\mathbf{X})$ – preço da habitação $\mathbf{X}$
$\mathbf{X} = (X_1, X_2, \ldots, X_n)$ – vector dos atributos determinantes na formação do preço da habitação
$\mathbf{b} = (b_1, b_2, \ldots, b_n)$ – vector dos *preços hedónicos* de cada atributo $X_{i\ i=1,\ldots,n}$
$\boldsymbol{\varepsilon}$ - componente estocástica

Para além da simplicidade, é necessário ter em conta que, ao contrário de outras formulações, o modelo de regressão linear implica que os *preços hedónicos* marginais, associados aos atributos, sejam constantes.

[5] A transformação Box-Cox, foi introduzida por Box G., & Cox D., (1964) como uma proposta para resolver o problema de estimação em regressões não lineares.

II.4.3. Variáveis independentes de um modelo hedónico

O estudo dos atributos físicos e estruturais de uma habitação são os mais utilizados em modelos de determinantes do preço da habitação. A área, o número de quartos, a idade, entre outros são características da habitação a que empiricamente se associa uma relação com o preço. Contudo, estes são apenas parte dos aspectos que determinam o valor de uma habitação. Como vimos anteriormente, várias teorias foram propostas para o mecanismo de formação de preços que apontam factores como a acessibilidade, a vizinhança, aspectos sociais, entre muitos outros.

A teoria subjacente à construção de *modelos de preços hedónicos* e os diversos trabalhos de investigação que têm vindo a ser desenvolvidos, não apontam um conjunto inequívoco de atributos que devem entregar os modelos. Designadamente, os exemplos apontados por Baranzini et al (2008), demonstram que as atenções comuns nos investigadores, direccionam-se para conjuntos restritos de atributos, dentro das temáticas que pretendem estudar. O mesmo autor aponta exemplos de diferentes temáticas de investigação que vêm sendo desenvolvidas:

- Influência de aspectos socioeconómicos: abordam-se, por exemplo, estudos relativos à estrutura da população residente na envolvente de uma habitação.
- Influência (positiva ou negativa) de factores ambientais: utilizados para a avaliação de impactos ambientais expectáveis e de externalidades ambientais. Constitui uma metodologia alternativa às abordagens de "*preferências declaradas"*.
- Acessibilidade territorial: avaliação quantitativa associada ao valor de centralidades genéricas, representadas por concepções socioeconómicas como é exemplo o CBD (*Central Business District*). Assume grande relevância ao nível local pela possibilidade de quantificar a acessibilidade a equipamentos e serviços.

Malpezzi (2008) refere que a capacidade explicativa de um *modelo hedónico* é dependente da utilização correcta de um leque de atributos, que represente as principais características de mercado da habitação. O mesmo autor alerta que a não consideração de atributos chave introduz enviesamento na estimação dos *preços hedónicos* das

variáveis consideradas. Tal como demonstrado por Linneman (1980), é frequente que os valores dos coeficientes nestes modelos sejam subestimados ou sobrestimados.

Para evitar estes possíveis desajustamentos, Malpezzi (2008) aponta a necessidade de o conjunto de atributos $X_{i,\ i=1,2,\ldots,n}$ referir três temáticas fundamentais, tal que:

$$p(\mathbf{X}) = f(\mathbf{E}, \mathbf{V}, \mathbf{L}) \qquad [3]$$

sendo:

$p\ (\mathbf{X})$ – preço da habitação;
X – habitação representado pelo triple (**E**, **V**, **L**), com:
E - atributos estruturais da habitação
V - características de vizinhança
L - atributos de localização (espacial)

Existe uma limitação associada aos pressupostos do *modelo hedónico*, que traz importantes desafios relativos à inclusão de atributos chave no modelo. Como vimos anteriormente, pelas características do bem habitação, o pressuposto de um mercado único pode não ser plausível. Um exemplo pode ser dado pelas transacções de compra e arrendamento. É expectável que cada tipo de contracto, associado ao tipo de transacção, apresente custos de transacção próprios. São sujeitos a exigências administrativas e fiscais diferenciadas, bem como são sensíveis a condições de financiamento específicas, resultando em dinâmicas autónomas.

Malpezzi (2008) refere duas alternativas para o tratamento deste problema:

i) especificar, no *modelo hedónico*, um conjunto de atributos que permitam distinguir ambas as habitações;

ii) tratar estas diferenças substanciais como submercados, e optar pela sua análise autónoma;

A última opção tem sido abordada em muitos trabalhos uma vez que a identificação e análise de submercados pode constituir um aspecto importante no conhecimento das dinâmicas habitacionais e dos mecanismos de mercado subjacentes. Facilmente se depreende que esta abordagem enfrenta um novo desafio, aqui ainda não referido: a identificação e delimitação de submercados – que não serão abordadas neste trabalho.

É previsível que os atributos de uma habitação tenham uma natureza quantitativa. Contudo, tal como noutros problemas econométricos, a definição de atributos pode

implicar a necessidade de englobar no modelo parâmetros qualitativos. Na realidade estes são frequentemente mais simples de obter pelo que têm de ser apresentadas algumas considerações sobre a utilização deste tipo de atributos.

A introdução de parâmetros qualitativos deve-se à necessidade de avaliar atributos que não são facilmente quantificáveis. Como exemplo, destacam-se trabalhos que pretendem avaliar o impacto de estigmas associados a espaços urbanos (como os bairros sociais ou outras áreas com óbvias segregações sociais. A forma alternativa de avaliar estes aspectos passa pela utilização de uma variável de tipo binário. A codificação, pode ser obtida, por exemplo para o caso de uma habitação localizada num determinado bairro, como dentro da área territorial (1) ou fora (0).

A utilização deste tipo de atributos (do tipo *Sim ou Não*) é comummente designada em econometria por *variável dummy*. Dependendo do tipo de modelo considerado e sendo uma variável de tipologia diferente, esta deve ser alvo de maior cuidado interpretativo. Wooldridge (2005) sugere que num modelo de regressão linear como o apresentado em [2], um atributo *dummy* pode ser tratado de forma semelhante aos restantes sempre que o seu preço (*hedónico*) seja determinado a partir de um método comum, como os mínimos desvios quadrados. O autor alerta que a consideração de um dado conceito sob uma *variável dummy* deve acautelar a inexistência de colinariedade. Para ilustrar esta questão, é dado o exemplo da utilização da variável género. Neste caso, a utilização de uma *variável dummy* para o género masculino e outra para o género feminino é contraproducente, uma vez que são variáveis mutuamente exclusivas e, obviamente, é expectável uma relação linear perfeita entre ambas as variáveis, introduzindo um erro grave no modelo. Em muito casos, este tipo de variáveis reflecte escolhas individuais ou outras unidades económicas e não atributos pré-determinados, como o género. A necessidade de avaliar a casualidade dessa informação torna-se ainda mais importante.

Numa relação linear como a descrita em [2], a interpretação de uma *variável dummy* é um pouco diferente dos atributos quantitativos. Aplicando uma ferramenta que permita determinar a relação linear e os preços (*hedónicos*) de um dado atributo, os coeficientes *b* indicam-nos o valor por unidade do atributo da habitação. Por outro lado, na *variável dummy*, o coeficiente deve ser lido como o acréscimo monetário associado ao atributo de referência, quando todos os outros atributos da habitação se mantêm constantes.

II.2.4. Limitações dos modelos de preços hedónicos

As características intrínsecas ao funcionamento do mercado da habitação apontam para a existência de limitações associadas à aplicação dos princípios de *modelos hedónicos*.

Equilíbrio de mercado

Sendo a teoria de *Lancaster* baseada no pressuposto de um mercado de concorrência perfeita em equilíbrio é necessário relembrar que a habitação tem associados fenómenos específicos que a afastam destas condições.

No entanto, no curto prazo, o *stock* existente domina o mercado, o que torna razoável assumir que a oferta de habitação é fixa, traduzindo-se em duas considerações:

i) o mecanismo de formação de preços entra facilmente em equilíbrio e é determinado, essencialmente, pela procura.

ii) no equilíbrio, os preços são razoavelmente estáveis ao longo do tempo.

Definição e mensuração de atributos

Um aspecto importante é a impossibilidade de estabelecer e quantificar, de forma unívoca, todos os atributos que determinam o preço da habitação. Tal como já mencionado, este facto influencia a capacidade explicativa do modelo e introduz um grau de incerteza na estimação dos coeficientes.

Estas dificuldades são óbvias na definição e mensuração de atributos com dimensão espacial. Sobressaem os aspectos relacionados com a localização da habitação. As condições de vizinhança são definidas por múltiplas características. Estas representam configurações diferentes para cada consumidor, não envolvendo, muitas vezes, o mesmo conjunto de atributos.

A dimensão espacial vem sendo incorporada numa grande diversidade de trabalhos, envolvendo variadas propostas e tipos de aplicações. Os trabalhos de Jud (1980), de Adair et al (2000) ou de Marques et al (2010) apontam diversas alternativas para a incorporação da localização relativa de cada habitação. Realçam-se as opções

metodológicas que recorrem à utilização de variáveis do tipo *dummy*, de medidas de distância ou ainda à aplicação do conceito de potencial[6].

Não são claras as vantagens e desvantagens de cada abordagem, cabendo ao investigador delimitar as potencialidades e constrangimentos de cada técnica para a concretização dos objectivos propostos. Embora a sua utilização levante questões essencialmente teóricas também é verdade que autores como Ross et al (2009) apontam que muitas das abordagens provocam inconsistências nos *modelos hedónicos* construídos.

Definição e mensuração do indicador preço

A frequente dificuldade em obter o preço de transacção verdadeiro coloca desafios à correcta selecção de uma variável que exprima o preço da habitação. Este pode ser interpretado como um conceito maleável, não sendo definido sempre da mesma forma. Por exemplo, é comum na maior parte dos estudos a utilização de um indicador de preço que é estimado por um profissional, por exemplo, de mediação imobiliária.

Malpezzi (2008) comenta que alguns trabalhos de investigação que se têm debruçado sobre esta problemática referem que a variância associada a estes indicadores é suficientemente elevada para permitir a aplicação de um *modelo hedónico* com segurança. Centrando-se na mensuração das diferenças do valor da habitação para diferentes conjuntos de atributos é expectável que o valor relativo de um dado atributo seja semelhante, independentemente dos indicadores de valor considerado. O mesmo autor, alerta para alguns estudos recentes que começam a questionar este pressuposto, embora ainda não de forma generalizada.

[6] O conceito de potencial é inspirado nas leis da física propostas por Newton. Encontra-se uma breve introdução à utilização deste conceito no contexto do estudo do espaço económico em Lopes (2001)

III. DATA MINING E ANÁLISE DE DADOS DO MERCADO IMOBILIÁRIO

A construção de um modelo de preços hedónicos envolve a utilização de dados produzidos nas transacções (ocorridas ou potenciais). Cada transacção tem dois tipos de informação essencial: o preço a que ocorre e as características do produto transaccionado.

Como é expectável, a distorções naturais dos mecanismos de mercado – que impossibilitam o seu funcionamento perfeito, tornam a detenção de informação um aspecto importante. Naturalmente não é facilmente disponibilizada, ocorrendo as transacções sob um certo manto de confidencialidade.

Neste campo, o poder de cruzar e utilizar todos os tipos de informação torna-se absolutamente relevante. Na utilização de técnicas de análise de dados tradicionais, a recolha de dados era promovida pelo próprio investigador. A maior parte das vezes, dependente da disponibilização de dados de domínio público ou, alternativamente adquirindo os seus próprios dados (quer por mecanismos de produção próprios – como a observação, quer adquirindo esses mesmos dados a instituições privadas).

A existência de cada vez maiores volumes de dados de domínio público (ou quase público), proporcionada pelo desenvolvimento dos meios de comunicação e informação, trouxe novos desafios. Como extrair o conhecimento dessas novas bases de dados? É a resposta a esta questão que procuramos desenvolver neste capítulo, proporcionando um enquadramento geral da metodologia de análise de dados seleccionada e da forma como esta se encaixa na resposta aos desafios colocados pela aplicação do modelo econométrico de preços hedónicos.

III.1 – Enquadramento

III.1.1. O Data mining e a análise de dados

O explosivo crescimento do volume de dados produzidos nas últimas décadas tornou essencial o desenvolvimento de novas técnicas e ferramentas capazes de analisar grandes volumes de dados e transformar a informação que contêm em conhecimento útil. Fayyad et al (1996) reflectem sobre este fenómeno e propõe uma definição simples que descreve o conceito de *data mining* como o *processo não trivial de extrair conhecimento útil e compreensível, previamente desconhecido, de grandes volumes de dados*.

A necessidade de desenvolver um processo automático ou semiautomático para a extracção de conhecimento a partir de bases de dados é uma resposta a cinco grandes desafios, referidos por Tan et al (2006), que motivaram o desenvolvimento do *data mining* em detrimento à utilização das metodologias de análise de dados tradicionais:

i) Dimensionalidade: o crescimento exponencial da capacidade de processamento e armazenagem dos sistemas informáticos e consequente decréscimo dos custos de geração, recolha e armazenamento de informação resultam em novas e multifacetadas estruturas de bases de dados com uma vastidão de atributos associáveis a cada objecto de estudo.

ii) Escalabilidade**:** as técnicas e algoritmos desenvolvidos devem ser altamente escaláveis, ou seja, devem funcionar com grandes volumes de dados (tera-bytes de informação). Como norma, as técnicas de análise de dados tradicionais podem funcionar bem com pequenas amostras mas não ser o suficientemente escalável para tratar bases de dados enormes.

iii) Heterogeneidade e complexidade: as bases de dados armazenam novas tipologias de atributos diferentes aos atributos tradicionais (categóricos e numéricos) comummente utilizados em análise de dados. Por outro lado, têm surgido novas e complexas fontes de dados provenientes de páginas webs, redes sociais, fluxo constante de dados (data streams), sistemas de informação geográfica, etc.

iv) Distribuição da informação**:** a circulação da informação é um aspecto fulcral na capacidade de aceder, armazenar e disponibilizar informação. Torna-se essencial responder a desafios como: reduzir a quantidade de informação, consolidar os resultados da análise de dados provenientes de múltiplas fontes, e garantir a segurança da informação, do ponto de vista da sua qualidade e também da privacidade.

v) Paradigma de análise: o princípio tradicional das técnicas de análise de dados baseia-se na utilização de testes de hipótese, a partir do qual se desenha um dispositivo experimental que permite recolher os dados necessários para o teste e posterior análise da hipótese colocada. A dimensionalidade, heterogeneidade e complexidade da informação disponível torna extremamente complexo este processo, que obrigaria, por exemplo, à colocação de centenas de hipóteses.

É de referir que o *data mining* surge da convergência de várias disciplinas, nomeadamente, a estatística clássica, a inteligência artificial e a aprendizagem automática, entre outras, para responder a todos estes desafios colocados pela rápida evolução dos recursos e tecnologias computacionais verificados nas últimas décadas.

III.1.2. Aplicações ao mercado imobiliário

Podemos encontrar a informação sobre os atributos de uma habitação em bases de dados constituídas com objectivos comerciais muito distintos. Como exemplos de agregadores de informação imobiliária podemos referir os promotores e construtores, os proprietários e os mediadores imobiliários. Podemos ainda encontrar dados acumulados nas estruturas da administração pública. Estas entidades recolhem, produzem e agregam dados. Dispõem de um potencial de informação útil ainda não identificada e extraída.

No caso do mercado imobiliário, a produção de nova e sistematizada informação é um elemento muito importante. Como referido anteriormente, a escassez de informação e da sua disseminação por todos os agentes resulta em diversas imperfeições de mercado que implica distorções ao mecanismo de formação de preços.

Exemplos das vantagens da utilização desta informação podem ser encontrados em portais imobiliários como o *Zillow*[7]. Uma parte importante dos inúmeros serviços disponibilizados neste website tem por base processos semi-automatizados de extracção de conhecimento a partir de dados. Podemos, como exemplo, referir a ferramenta *Zestimate*. A construção deste índice de preços da habitação permite ao utilizador obter uma estimativa do valor de mercado de qualquer habitação cadastrada numa área territorial, esteja a referida habitação no mercado ou não. A ferramenta cruza múltiplas fontes de informação, destacando-se a utilização de bases de dados que permitem extrair o valor dos atributos de habitações transaccionadas no passado recente. Utilizando as informações cadastrais que permitem, entre outras coisas, identificar os principais atributos das habitações existentes numa dada unidade territorial, é possível criar estimativas do preço para todas as habitações existentes.

Por outro lado, portais de promoção imobiliária como o portal Nacional de Imobiliária Casa Sapo[8] acumulam informação semelhante à utilizada nesta ferramenta. Os serviços do portal ainda não fornecem ferramentas de índice de preços como o *Zestimate*, desperdiçando conhecimento que possa estar contido nos volumes de dados que acumula. A disponibilização de parte da base de dados deste portal permite o desenvolvimento, neste trabalho, de uma primeira aproximação às vantagens da utilização das técnicas de *data mining* para extracção de conhecimento.

Dada a múltipla diversidade de indicadores e de informação associada ao mercado imobiliário e à habitação em particular, disponibilizada nas bases de dados, um dos desafios mais comuns colocados aos investigadores quando pretendem utilizar abordagens baseadas em modelos de *preços hedónicos* centra-se precisamente na necessidade de identificar os atributos determinantes no preço da habitação. Este é um ponto de partida incontornável e traduz-se essencialmente num problema de selecção correcta dos indicadores disponíveis.

A importância da tarefa de selecção de indicadores (atributos) relevantes na concretização dos objectivos deste trabalho traduz o valor associável à implementação de um processo de *data mining*. Este processo assume-se como uma alternativa às tradicionais metodologias de análise de dados, onde as tarefas de selecção são muito dependentes da capacidade técnica do investigador.

---

[7] *http://www.zillow.com/*
[8] *http://www.casa.sapo.pt/*

III.2 – Desenho de um processo de data mining

III.2.1. Metodologia padrão

A existência, nas bases de dados, de informação irrelevante para o problema em investigação, torna a análise de um dado problema num processo complexo e moroso. O problema é ainda maior quando pretendemos utilizar dados que não foram efectivamente recolhidos com esse objectivo. A abordagem metodológica proposta para a utilização de *data mining* permite agilizar o planeamento do processo global de análise, fornecendo um conjunto de ferramentas, implementadas de forma iterativa, onde ressalta a necessidade de sucessivas repetições de tarefas. A Figura 2 procura apresentar a visão geral, consensual na bibliografia consultada, relativa a um qualquer processo de extracção de conhecimento a partir de uma base de dados.

Figura 2 Processo geral de extracção de conhecimento de uma base de dados
(fonte: http://www.fp2.com.br/datamining/?cat=3)

A implementação de ferramentas de *data mining* tem-se disseminado por várias actividades comerciais e industriais. O trabalho desenvolvido pelo grupo CRISP-DM[9] (CRoss-Industry Standard Process for Data Mining) é um exemplo da importância e necessidade de desenvolver um padrão metodológico.

Envolvendo diversos actores, onde se incluem empresas que desenvolvem soluções de análise de dados (a SPSS), entidades industriais como a Daimler Chrysler ou ainda o financiamento da União Europeia, o grupo reuniu esforços com o objectivo de desenvolver um método de implementação padrão para as tarefas de *data mining*, independente do *software* e do sector económico em que se desenvolve.

[9] CRISP-DM é um projecto colaborativo de empresas como a Teradata, a SPSS, a Daimler Chrysler, a NCR e a OHRA. Mais informação encontra-se disponível em: http://www.crisp-dm.org/

Baseado numa abordagem do tipo "*ciclo de vida"*, o processo iterativo e adaptativo, baseia-se na repetição do ciclo de tarefas para aprimorar e aproximar a solução final do objectivo proposto.

O ciclo (Figura 3) engloba seis fases essenciais. Cada fase é composta por diversas tarefas, descritas de forma hierárquica por níveis de abstracção, que permitem tornar o processo perceptível a investigadores com *backgrounds* diferenciados.

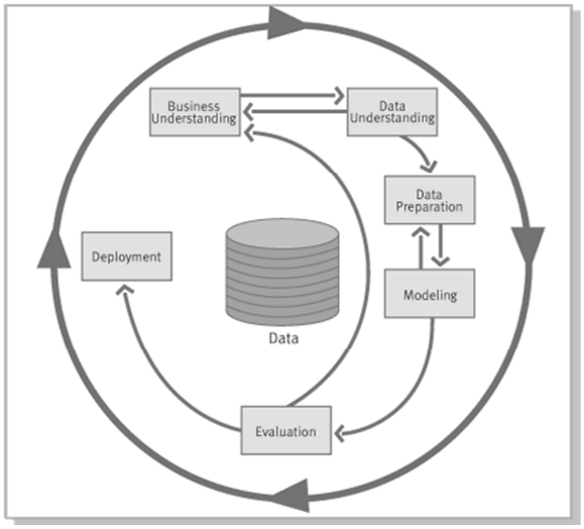

Figura 3 Processo de extracção de conhecimento proposto pelo grupo CRISP-DM (fonte: CRISP-DM group)

As fases descritas na Figura 3 referem-se a:

Entendimento do negócio *(Business understanding):* envolve a definição de objectivos e requisitos de um projecto e da necessidade de implementar um processo de *data mining* dentro de uma dada organização. Baseado no conhecimento adquirido, esta fase abarca ainda o desenho do tipo de tarefas a implementar.

Entendimento dos dados (*Data understanding*): a recolha dos dados envolve não só os mecanismos de acesso como uma avaliação descritiva das suas características e dos níveis de qualidade (despistando os problemas associados a erros e omissões, por exemplo). Assim é possível estabelecer um primeiro conjunto de hipóteses bem como detectar conjuntos de variáveis com especial interesse para o tema em estudo.

Preparação dos dados (*Data preparation*): inclui a execução de todas as tarefas e respectivas ferramentas que permitem construir um conjunto de dados que responda aos requisitos da fase seguinte, de modelação. Incluem-se tarefas como a limpeza, a transformação dos dados e a selecção (de subconjuntos de variáveis, de variáveis

individuais, de objectos). O processo é repetido múltiplas vezes e integrado de forma iterativa com a fase de modelação.

Modelação (*Modelling*): para um dado problema ou hipótese são seleccionadas diversas ferramentas e modelos, das variadas opções existentes. Embora dedicadas à resolução do mesmo tipo de problemas, cada técnica encerra em si requisitos próprios que obrigam por um lado ao investigador a optimizar os seus parâmetros para os objectivos estabelecidos e por outro, a garantir requisitos específicos nos dados de entrada.

Avaliação (*Evaluation*): o(s) modelo(s) construídos devem ser avaliados com correctas medidas de qualidade e também com a revisão de todos os passos e opções que foram sendo tomadas durante a execução de todo o processo. O investigador deverá certificar-se que o modelo representa uma boa solução para os desafios que tinham sido colocados ao processo de *data mining* e deverá estar em condições de decidir sobre a validade da utilização dos resultados.

Implementação (*Deployment*): consideradas boas soluções, os modelos obtidos são incorporados em processos de tomada de decisão e em novos estudos analíticos. Esta fase representa a integração dos resultados nas actividades da organização.

As tarefas incluídas em cada uma das fases anteriores são descritas no guia CRISP-DM por níveis de abstracção. A ideia é que facilmente se possa desenhar todo o processo, sem necessidade de desenvolver previamente um conhecimento aprofundado de cada uma das ferramentas existentes.

Estes quatro níveis de abstracção são:

i) Um nível de topo, que se refere à necessária reflexão da pertinência da utilização de ferramentas de *data mining* para a concretização de objectivos de um trabalho de investigação. Envolve a delimitação dos objectivos e o estudo teórico da problemática bem como a análise e descrição das bases de dados disponibilizadas.

ii) O segundo nível hierárquico aponta para as tarefas genéricas associadas a cada fase. Engloba, por exemplo, a selecção de processos de recolha e descrição dos

dados, a implementação de regras de selecção de dados, a escolha de modelos a utilizar, o desenho do processo de avaliação, entre outras.

iii) As tarefas genéricas anteriores são estabelecidas a um terceiro nível por tarefas especializadas, com especificação das abordagens convenientes para diferentes situações práticas. Por exemplo, a escolha de métodos de selecção de atributos do tipo filtro ou embutida, a implementação do método de validação *k-fold* ou *holdout*, entre outros.

iv) O último nível refere-se à escolha dos operadores disponíveis dentro de uma dada solução computacional. Neste nível de abstracção exige-se o domínio do investigador de cada um dos algoritmos que selecciona e a capacidade de os manipular para a utilização numa dada tarefa especializada.

III.2.2. Implementação do processo

Ajustado aos objectivos propostos e às limitações dos dados disponíveis, apresenta-se o desenho do processo de *data mining* implementado neste trabalho. Para cada fase definiu-se um conjunto de tarefas e respectivos métodos e técnicas seleccionadas.

A exploração das técnicas e ferramentas de *data mining* no estudo de uma problemática ligada ao planeamento do território constitui o pano de fundo deste trabalho, a que não são alheios os requisitos pedagógicos associados ao desenvolvimento de uma dissertação de mestrado. Assim, optou-se por não desenvolver detalhadamente todas as fases metodológicas do modelo CRISP–DM, uma vez que este é orientado para o desenvolvimento de projectos em ambiente profissional e empresarial. Nos parágrafos seguintes procura-se descrever as adaptações implementadas (Figura 4).

CRISP - DM	Metodologia de *data mining* implementada
Entendimento do negócio	**Definição de objectivos**
Entendimento dos dados	**Recolha de dados**
Preparação dos dados	**Pré - processamento**
Modelação	**Modelação**
Avaliação	**Avaliação**
Aplicação	

Figura 4 Fases do processo de *data mining* implementado

1. Definição de objectivos

A identificação dos determinantes no preço da habitação a partir de um modelo de *preços hedónicos* é o problema de investigação para o qual se estabelece a necessidade de desenhar um processo de análise de dados.

A implementação do processo de *data mining* enquadra-se na resposta a este objectivo uma vez que o trabalho é caracterizado por dificuldades na recolha de dados e respectiva selecção dos correctos indicadores que nos permitem responder ao problema investigado.

2. Recolha de dados

A recolha dos dados exige a inventariação dos recursos disponíveis que possam fornecer indicadores para a tarefa de investigação, assim como analisar a capacidade de aceder a essas fontes de informação.

Identificados os recursos existentes e garantida a capacidade (ou autorização) de aceder às bases de dados, torna-se necessário definir o processo de importação de dados para que englobe os requerimentos necessários às tarefas do processo desenhado. Efectua-se uma avaliação dos constrangimentos que os indicadores apresentem e desenham-se novas tarefas, a incluir nas fases seguintes do processo.

Inclui-se nesta fase a descrição sumária das características mais comuns dos indicadores utilizados: tipo de variável, unidades de medida, estatísticas descritivas (*média*, *desvio padrão*, *máximos*, *mínimos*), entre outras que se considerem relevantes.

3. Pré - Processamento

O volume de dados recolhido exige frequentes operações de limpeza e transformação que permitam reconstruir a base de dados de forma a cumprir todos os requisitos necessários para a execução dos métodos e técnicas a serem implementadas. Quando os dados são demasiado complexos (elevado número de atributos) ou em volume elevado (elevado número de registos), a exploração prévia de dados torna custosa a implementação das ferramentas de pré-processamento. O *data mining* permite conjugar a concretização de tarefas tradicionais com implementações computacionais semi-automatizadas que facilitam e agilizam este processo. Genericamente, a conjugação de múltiplas estratégias e técnicas de pré processamento devem centrar-se no objectivo de melhoria do desempenho dos algoritmos de modelação dos dados.

O guia metodológico CRISP-DM descreve como tarefas principais da preparação de dados a:

- Integração – combinação de diferentes bases de dados
- Limpeza – eliminação de inconsistências, tratamento de valores omissos e estruturação dos dados nos formatos requeridos pelos métodos a utilizar;
- Formatação dos dados – modificações que melhorem o desempenho dos algoritmos e técnicas na fase de modelação;
- Construção – produção de atributos derivados dos dados iniciais;
- Selecção – selecção dos atributos relevantes para a construção do modelo que permita descrever os conceitos inerentes à problemática;

A selecção de atributos ganha especial relevância dentro do conjunto das tarefas de pré–processamento previamente descritas, uma vez que, o desenho deste processo, tem como objectivo essencial o problema da identificação de atributos valorativos do bem habitação. Dada esta importância, apresenta-se na secção III.3 uma descrição detalhada das diferentes tarefas de selecção de atributos implementadas.

4. Modelação

O modelo econométrico de *preços hedónicos* é estabelecido a partir da determinação da relação entre uma variável dependente, que representa o preço de um bem, e um conjunto de variáveis independentes, que se referem a atributos desse mesmo

bem, sendo a relação linear um pressuposto plausível no contexto deste trabalho, tal como proposto na secção II.4.2, equação [2].

Os modelos de regressão são as ferramentas óbvias para determinar este tipo de relação, destacando-se a regressão linear multivariada que, sendo a única ferramenta utilizada na fase de modelação, será analisada em maior pormenor na secção III.4.

5. Avaliação

A fase de avaliação tem como função uma análise crítica e preliminar dos resultados do modelo implementado. Pretende-se determinar a coerência e consistência dos modelos para a concretização dos objectivos propostos. Constitui ainda a oportunidade de reavaliar todo o processo, podendo levar à sua repetição com modificação ou alteração de tarefas ou ferramentas utilizadas.

A avaliação requer a selecção de *medidas de desempenho* adequadas. No problema de modelação baseado na ferramenta de regressão linear multivariada ressaltam duas medidas de avaliação comummente utilizadas em ciências sociais:

i) Estatística de teste para a hipótese da efectiva relação entre variável dependente e variáveis independentes, dada pelo teste F.

ii) Coeficiente de medição da capacidade explicativa do modelo face à estrutura dos dados utilizados.

O valor de uma *medida de desempenho* pode ser estimado a partir de um esquema de validação seleccionado. Tradicionalmente, o mesmo conjunto de dados disponível é utilizado para construir o modelo e estimar as medidas de desempenho. Porém, nos volumes grandes de dados o custo computacional pode torna-se incomportável. É necessário realçar que a utilização do mesmo conjunto de dados para construir e testar o modelo produz estimativas muito optimistas do desempenho. Em geral, a abordagem correcta passa por avaliar tendo em conta a capacidade de generalização aos dados que não foram usados para construir o modelo.

Com este objectivo foram desenvolvidas abordagens alternativas de validação baseadas em técnicas de amostragem. A ideia mais simples consiste em particionar o conjunto de dados em dois: um *conjunto de treino* com os exemplos que são usados para construir o modelo e um *conjunto de teste* com os exemplos que são usados para estimar a medida de desempenho. Existem vários métodos de validação baseados em

diferentes partições do conjunto de dados. Estes métodos, conjuntamente com as medidas de desempenho utilizadas neste processo de *data mining,* são descritos pormenorizadamente na secção IV.5 deste capítulo.

III.3 – Selecção de atributos

O *data mining* compreende um conjunto de ferramentas de selecção de atributos que permitem agilizar esta tarefa. Tan et al (2006) refere esta como uma tarefa essencial, uma vez que:

- ✓ os algoritmos de aprendizagem têm desempenhos superiores em bases de dados de dimensões reduzidas, onde atributos redundantes e irrelevantes para a construção do modelo são eliminados;
- ✓ permite maior simplicidade na análise dos modelos construídos;
- ✓ a quantidade de tempo, memória e processamento necessários para os algoritmos de modelação são menores em conjuntos reduzidos de variáveis.

A importância desta tarefa para a concretização dos objectivos propostos impõe uma explanação mais detalhada das técnicas de selecção implementadas. Podemos referir três tipos de abordagens, que englobam critérios supervisionados (medidos em função de uma variável resposta) e não supervisionados (que envolvem outro tipo de critérios).

1. *Redução da dimensionalidade:* reduz a dimensionalidade dos dados iniciais de forma não supervisionada, envolvendo a transformação dos dados em novos atributos a partir de técnicas fundamentalmente baseadas em álgebra linear.
2. *FSS (feature subset selection*): permite seleccionar a partir de todos os atributos iniciais um subconjunto para posterior análise. A ideia genérica consiste em testar todos os possíveis subconjuntos de atributos e seleccionar aquele que obtém melhores resultados. Na impossibilidade de levar a cabo esta estratégia de busca exaustiva no espaço de possíveis subconjuntos devido ao grande custo computacional que a sua implementação envolve (para um conjunto de n atributos devem ser avaliados 2^n subconjuntos), foram desenvolvidos três tipos de abordagens que podem ser implementadas:
    i) *selecção por filtro*: as técnicas de selecção são implementadas previamente ao processo de modelação; recorre a critérios de selecção supervisionados ou não supervisionados

ii) *selecção por wrapper*: semelhante à abordagem filtro mas utilizando a medida de desempenho do próprio algoritmo de aprendizagem durante o processo de modelação como critério de selecção. É comummente classificada como uma abordagem supervisionada

iii) *abordagem embutida*: a selecção de atributos é efectuada como parte do algoritmo de aprendizagem durante o processo de modelação, podendo utilizar critérios supervisionados ou não supervisionados

3. Uma alternativa complementar que permite seleccionar um conjunto restrito de variáveis, é a utilização de um *esquema de pesagem* a partir de uma ferramenta que permita descrever e hierarquizar as estruturas presentes nos dados.

Com excepção da abordagem de selecção *wrapper*, não utilizada neste trabalho por requerer um esquema de implementação mais complexo e apresentar um elevado custo computacional, foram implementados todos os restantes métodos de selecção atrás descritos e que se descrevem de forma mais detalhada nas subsecções seguintes.

III.3.1. Redução da dimensionalidade

Das técnicas de redução da dimensionalidade apresenta-se a Análise de Componentes Principais (ACP) por ser uma ferramenta apreendida no âmbito do percurso académico em Planeamento Regional e Urbano. Esta técnica tem como objectivo encontrar uma transformação mais representativa e geralmente mais compacta das observações.

Com este fim, o método de ACP aplica uma transformação linear para projectar o vector de variáveis iniciais $\mathbf{X} \in \Re^n$ noutro vector $\mathbf{Y} \in \Re^m$ (para $m \leq n$). As componentes do vector **Y** são designadas *componentes principais*. Cada componente principal é uma combinação linear de todas as variáveis originais. Estas componentes são independentes entre si e individualmente responsáveis pela variância das observações. Como a parte mais importante da variância dos dados iniciais é explicada por um número reduzido de componentes, torna-se possível descartar as restantes, sem perdas importantes de conhecimento e reduzindo a informação explicativa do problema.

O método é construído através da medição das relações existentes nas variáveis observáveis de **X** determinada a partir de uma matriz de covariâncias ou de uma matriz de correlações. Os valores próprios e respectivos vectores próprios, calculados para uma destas matrizes, permitem definir uma transformação linear dos dados iniciais. As componentes principais são definidas pelos vectores próprios: a primeira componente principal é o vector próprio associado ao valor próprio mais elevado, a segunda componente é o vector próprio associado ao segundo valor próprio mais elevado, etc.

Os valores próprios podem ser tomados como indicadores da variância explicada dos dados iniciais. Tal como se ilustra na Figura 5, existem componentes que variam relativamente pouco comparativamente com outras (é o caso das componentes associadas aos valores próprios mais pequenos). Este facto permite descartar, com segurança, componentes de menor variância sem afectar substancialmente a qualidade dos dados e consequentemente a sua interpretação.

A selecção do número de componentes que permitem uma descrição dos conceitos latentes nos dados iniciais é realizada tendo em consideração:

- para um número de componentes no modelo final menor que 30: utilização do critério de Kaiser, que se baseia na escolha de componentes cujos valores próprios são superiores a 1. Este critério corresponde maioritariamente das vezes à selecção de um número de componentes que permite cumulativamente a explicação de aproximadamente 75% da variância das variáveis iniciais.
- para um valor de componentes superior a 30: análise gráfica da variância cumulativamente explicada e do número de componentes, escolhendo os pontos no maior declive da curva. Quando o número de casos é efectivamente muito elevado, ambas as alternativas conduzem ao mesmo resultado.

Retido o conjunto de componentes principais que explicam a parte mais significativa da variância dos atributos iniciais, a caracterização dos conceitos subjacentes a cada componente é feita a partir dos pesos, também designados por *loadings* que definem a correlação de cada variável inicial com cada uma das componentes. Esta nova matriz permite ao investigador descrever, com um certo grau de subjectividade, o conceito que a componente caracteriza. Para ajudar nesta

interpretação é possível ainda determinar os *scores* – um valor estandardizado e unidimensional que aplica a transformação linear para cada um dos casos.

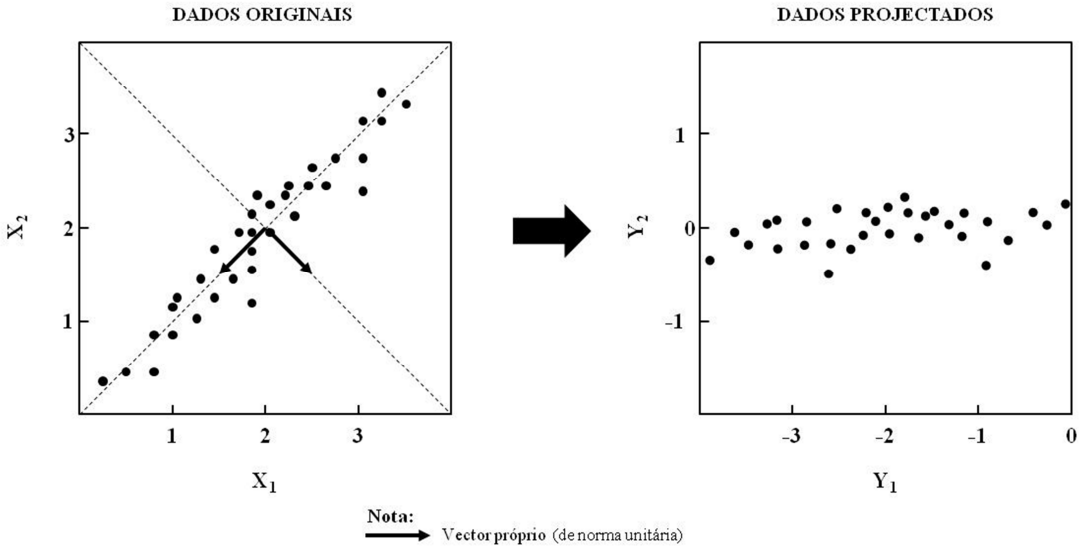

Figura 5 Os dados originais, altamente correlacionados e representados pelas variáveis $\mathbf{X_1}$ e $\mathbf{X_2}$ são projectados no novo sistema de coordenadas $\mathbf{Y_1}$ e $\mathbf{Y_2}$ (designadas componentes na ACP).
Como se pode verificar, com esta projecção as novas componentes apresentam variações muito díspares, ao contrário das variáveis iniciais; assim, é possível descartar a componente que traduz uma menor variação dos dados (neste exemplo, $\mathbf{Y_2}$) sem afectar significativamente a informação contida nos dados e, desta forma, reduzir a dimensionalidade de um dado problema.

III.3.2. Selecção de subconjuntos de atributos

Abordagem Filtro

A arquitectura genérica dos algoritmos de selecção de subconjuntos de atributos baseadas em filtros, baseia-se na definição e adopção de:

- Uma estratégia de busca;
- Critério de paragem;
- Medida de avaliação (supervisionada ou não – supervisionada)

Podemos identificar três tipos de estratégias de busca usualmente utilizadas: a estratégia *brute force* e duas estratégias que implementam uma heurística de busca greedy, *forward selection* (busca para a frente) ou *backward selection* (busca para trás).

A estratégia *brute force* procura testar todas as possíveis combinações de atributos. Como comentando anteriormente é extremamente custosa em termos computacionais. Esta pode só ser utilizada em processos muito específicos onde o número de variáveis a avaliar é substancialmente reduzido. A estratégias do tipo *greedy* introduz o critério direcção, designado *forward* se se referir à selecção através de um mecanismo de adição sucessiva de variáveis ou direcção *backward* quando se baseia na eliminação de variáveis a partir do conjunto inicial.

Para a abordagem filtro seleccionou-se apenas uma heurística de busca *greedy* com direcção *forward* e *backward*. No primeiro caso, a busca é iniciada para um conjunto sem atributos, sendo que estes são adicionados sucessivamente, seleccionando em cada passo, aquele atributo que melhor contribui para a medida de avaliação. No segundo caso, começa-se com todo o conjunto de atributos disponíveis e em cada passo, o algoritmo vai eliminando o atributo que menos contribui para a medida de avaliação. A medida de avaliação utilizada é a CFS – *Correlation-based feature selection*, descrita por Hall, A. & Smith, A., 1997. Esta medida avalia a efectiva importância de um subconjunto de atributos, dando relevância às variáveis altamente correlacionadas com a variável independente – o que torna o algoritmo de selecção supervisionado – e que sejam não correlacionadas com qualquer outra variável no conjunto.

Abordagem Embutida

A selecção de atributos pode ainda ser implementada em conjunto com o algoritmo de aprendizagem. O algoritmo de aprendizagem, do modelo de regressão linear, implementado no *software* disponibiliza dois algoritmos embutidos de selecção de atributos:

i) um algoritmo de busca *Greed*y com direcção forward, utilizando como medida de avaliação o critério de informação AIC (do inglês, *Akaike information criterion*) proposto em Akaike H., 1974. Esta medida é supervisionada e baseia-se no conceito de entropia. Tal como definido pela termodinâmica, de onde é originária, a entropia é uma medida de desordem de um sistema. Neste caso o conjunto de atributos seleccionados é aquele que minimiza o valor de *AIC* definido como $AIC = 2 * \frac{(n-L)}{N}$, onde *n* é o número de atributos, *N* o número de observações e L o logaritmo da verosimilhança, determinado como a

probabilidade de um conjunto de parâmetros θ de um modelo determinado para um conjunto de dados D, $L = P(\theta|D)$ – no caso particular aqui apresentado, de um modelo de regressão, temos $L = P(b|D)$, sendo b os coeficientes do modelo de regressão.

ii) indução de uma árvore de regressão utilizando o algoritmo M5prime. Tal como proposto por Wang Y. & Witten H. (1997), este algoritmo de aprendizagem supervisionado permite a construção de uma árvore de decisão. Este algoritmo sujeita os ramos da árvore obtida a um processo de poda recorrendo à construção de um modelo de regressão linear. Seleccionam-se desta forma os atributos mais importantes, em cada ramo, determinados pelos coeficientes do modelo de regressão.

Pesagem de atributos

As técnicas de pesagem permitem determinar para cada atributo X_i um peso w_i. Este peso reflecte a importância do atributo para a tarefa de modelação, ou seja, um atributo com maior peso será, à partida, mais relevante do que um atributo com menor peso. Os pesos de todos os atributos, ordenados, podem ser sujeitos a um processo de selecção que pode englobar a selecção dos *k* atributos com maiores pesos ou, alternativamente, aqueles com peso superior a um dado patamar.

Neste trabalho foram utilizados os resultados de dois métodos como base para a selecção por pesagem:

Pesagem utilizando ACP.

Os pesos dos atributos iniciais são estabelecidos pelos *loadings* de uma das componentes obtida no algoritmo de ACP (já descrito anteriormente), combinado com o critério de selecção. O critério mais comum baseia-se na utilização dos *loadings* da primeira componente principal, visto que é a componente que agrega maior capacidade explicativa da variância dos dados iniciais, utilizando-se como critério de selecção de variáveis um patamar de 0,500, que traduz a correlação mínima considerada do atributo inicial com a componente seleccionada.

Pesagem utilizando uma máquina de suporte vectorial

Desenvolvidas recentemente, as máquinas de suporte vectorial (Vapnik, 1998) podem ser utilizadas para a selecção de atributos através de técnicas de pesagem como proposto por Guyon (2002). A ideia em que se baseia a selecção é na utilização dos pesos w_i do classificador linear $Y = w\mathbf{X} + b$ resultante para obter uma lista ordenada dos atributos. O classificador linear é obtido como solução de um problema de optimização que procura determinar o hiperplano que melhor separa o conjunto de variáveis independentes iniciais. Os parâmetros $w_{i\ (i = 1, \ldots, n)}$, estandardizados, definem a correlação de um dado atributo com a fronteira de decisão classificatória deste modelo de dados hipotético, o que traduz a maior ou menor relevância do conceito subjacente a uma dada variável inicial (se a variável traduzir um conceito é expectável que contribua de forma mais acentuada para a definição da fronteira de decisão que traduz essa diferença de conceitos subjacentes nos dados iniciais). De forma semelhante à técnica de selecção pela ACP, a necessária implementação conjugada com um algoritmo de pesagem utiliza um patamar no valor de 0,500.

Podemos referir ainda uma abordagem híbrida entre a redução de dimensionalidade e as técnicas de selecção de atributos por pesagem. Definidas as componentes de uma ACP, por exemplo, os pesos mais elevados das variáveis iniciais para cada uma das componentes são as variáveis com um contributo mais importante para representar o conceito subjacente. Assim, ao invés de representar o problema a partir de novas variáveis (transformadas por combinação linear do conjunto inicial) é aceitável seleccionar estas variáveis como representativas dos conceitos subjacentes a cada componente. Assim, não teremos de analisar um modelo com variáveis dimensionais, muitas vezes difíceis de interpretar e transpor para a realidade.

III.4 – Regressão linear multivariada

A importância deste algoritmo para a modelação obriga-nos a apresentar alguns dos princípios básicos que estão subjacentes à técnica seleccionada que nos permite construir o *modelo hedónico*.

A relação matemática dada pela equação [2] pode ser obtida com a implementação de um algoritmo de regressão linear que determine a equação da recta que minimize o quadrado dos desvios dos valores efectivamente verificados.

De forma a simplificar a apresentação do algoritmo de regressão linear implementado, comecemos por tratar o problema de regressão linear simples.

A existência efectiva de uma relação linear entre duas variáveis pode ser medida através do cálculo do **coeficiente de correlação** linear. No caso simples de duas variáveis, com um conjunto n de pares de observações (x_i, y_i), $i = 1, \ldots, n$, o coeficiente de correlação linear é definido como:

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \textbf{[4]}$$

Como se pode verificar, a expressão anterior determina uma normalização do valor de R, pelo que o módulo de R nunca será superior a 1. Verifica-se ainda, que a relação pode ser positiva ou negativa, variando o coeficiente entre [-1, 1] e estando próximo dos extremos quando a relação entre as variáveis é efectivamente forte.

Sendo aceitável a relação do tipo linear, esta exprime-se em funções do tipo:

$$y_i = b_0 + b_1 x_{1,i} \text{ , com } i = 1, \ldots, n \qquad \textbf{[5]}$$

De forma a ilustrar o processo de cálculo dos parâmetros do modelo, a Figura 6 apresenta um exemplo de um conjunto de pontos (x_i, y_i) e de uma recta $y = b_o + b_1 x$ que representa um modelo probabilístico da relação linear estabelecida entre o conjunto de pontos.

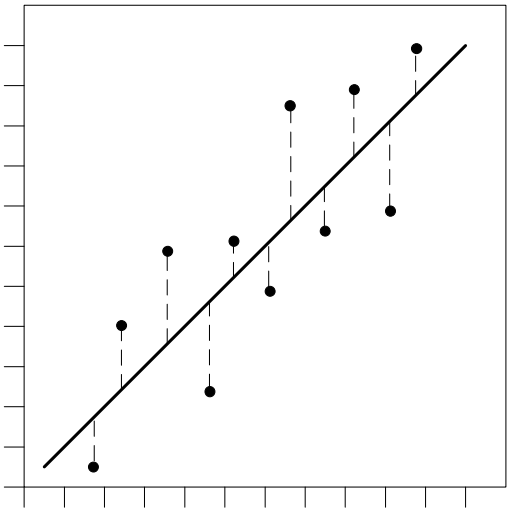

Figura 6 Modelo probabilístico da relação linear existente entre um conjunto de pontos.

Para assumirmos este como o melhor modelo é necessário verificar se os erros não são correlacionados. Alternativamente é possível testar se os erros são independentes (aleatórios), apresentando distribuição Normal, de média 0 e variância σ^2, uma vez que sob estas condições a independência é equivalente à não correlação.

Como já foi referido, sendo as variáveis independentes pré determinadas pelo investigador, assume-se o pressuposto de que estas sejam efectivamente independentes, e, dessa forma não são correlacionadas, independentemente da distribuição que tenham.

Satisfeitas estas condições, o método dos mínimos desvios quadrados é uma forma expedita de determinar os coeficientes de regressão. No problema de regressão linear simples, a solução que minimiza a soma dos quadrados dos desvios das observações y_i em relação aos seus valores estimados $\widehat{y_i}$,

sendo:

$$E[y_i|x_i] = \widehat{y_i} = b_0 + b_1 x_i \qquad [6]$$

e os desvios dados por:

$$y_i - E[y_i|x_i] = b_0 + b_1 x_i + \varepsilon_i - b_0 - b_1 x_i = \varepsilon_i \qquad [7]$$

com a soma de quadrados dos desvios definida como:

$$D = \sum_{i=1}^{n} {\varepsilon_i}^2 \qquad [8]$$

é a resolução do sistema que permite minimizar D tal que:

$$\begin{cases} \frac{\partial D}{\partial b_0} = 0 \\ \frac{\partial D}{\partial b_1} = 0 \end{cases} \Leftrightarrow \begin{cases} \hat{b}_0 = \hat{y} - \hat{b}_1 \overline{x} \\ \hat{b}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{cases} \qquad \textbf{[9]}$$

Como se pode verificar as expressões encontradas são bastante complexas para calcular manualmente. No caso de uma regressão linear múltipla, o volume de cálculos é ainda mais elevado, daí a dificuldade atrás referida da expansão dos *modelos de preços hedónicos* enquanto ferramenta econométrica.

Hoje, estes cálculos encontram-se implementados sob a forma de algoritmos computacionais em diversas soluções de software (incluindo em acessíveis "máquinas de calcular").

A recta de regressão obtida (para o exemplo de regressão linear simples atrás descrito) é dada por:

$$y = \hat{b}_0 + \hat{b}_1 x \qquad \textbf{[10]}$$

Assumidos os pressupostos anteriores, esta é melhor estimativa para a verdadeira recta $y = b_0 + b_1 x$. Os valores $\hat{y}_i$ obtidos a partir da equação [10] são designados comummente valores preditos.

De forma auxiliar é ainda comum o cálculo dos parâmetros *b* na forma *estandardizada* o que designaremos por β, e que são estimativas dos parâmetros quando as variáveis são sujeitas a uma operação de normalização prévia de média nula e variância unitária. Estes coeficientes indicam-nos a variação em desvios padrões, da variável dependente quando a variável independente varia um desvio padrão. No caso de um *modelo de preços hedónicos*, este parâmetro é útil de forma a hierarquizar comparativamente a importância da(s) variável(is) independente(s) no preço da habitação dado que a natureza diferenciada dos atributos corresponde a unidades de medida não comparáveis (por exemplo, a quantidade de atributo área é diferente da quantidade de atributo preservação ou acessibilidade).

Quando, a equação [5] representa uma equação de um um plano ($n = 2$) ou hiperplano ($n > 3$). Os cálculos são em tudo análogos aos apresentados para a regressão linear simples. Torna-se conveniente o tratamento matricial do problema de regressão

uma vez que simplifica os cálculos para este tipo de modelos pelo que se pode reescrever a equação [5] como:

$$\mathbf{Y} = \mathbf{X}\,\mathbf{b} + \boldsymbol{\varepsilon} \qquad [11]$$

Sendo e tal como apresentado em [3], para um *modelo de preços hedónico*s,

- ✓ **Y** - representado por **p** e refere-se a um vector (*n* x 1) de preços, de *n* habitações
- ✓ **X** - uma matriz (*n* x N).
- ✓ **b** - um vector (*n* x 1) de parâmetros de regressão, que, no modelo econométrico referem-se ao *preço hedónico* dos atributos **X**.
- ✓ **ε** - o vector (*n* x 1) de erros aleatórios.

III.5 – Métodos de validação e medidas de desempenho

Como referido anteriormente, existem vários métodos de validação baseados em diferentes partições do conjunto de dados que permitem estimar uma medida de desempenho.

Os métodos de validação implementados neste processo de *data mining* são dois:

Método de Validação Hold-Out.

Particiona a base de dados inicial num conjunto de treino, a partir do qual é construído o modelo, e outro conjunto de teste utilizado para estimar a medida de avaliação. As proporções de dados em cada partição são tipicamente decididas pelo analista, devendo ser tomadas algumas considerações:

a) utilização de um menor número de exemplos de treino poderá implicar uma menor capacidade de ajustamento do algoritmo de modelação ao conjunto de dados;

b) os subconjuntos não são independentes um do outro, podendo as partições resultar numa estrutura de dados diferenciada.

As partições comummente implementadas referem a utilização de 70% dos dados como conjunto de treino e 30% como conjunto de teste.

Método de Validação Cruzada k-fold

Esta técnica consiste na aplicação sucessiva de partições ao conjunto de dados, sendo por isso uma generalização do método hold-out. A ideia é dividir o conjunto de dados em k partições (subconjuntos) mutuamente exclusivos e de comprimentos aproximadamente iguais. O processo é repetido k vezes. Em cada iteração é seleccionada a actual k partição como conjunto de teste e as restantes k-1 partições como conjunto de treino. A estimação da medida de desempenho é dada pela média das medidas de desempenho decorrentes das k avaliações.

Na avaliação do desempenho de modelos de regressão linear por investigadores em ciências sociais é muitas vezes referida a utilização do coeficiente de determinação

R^2. Sendo uma medida bastante simples, fornece uma indicação da capacidade explicativa do modelo. De forma complementar, também é comum a utilização de um teste para a avaliação da relação de dependência entre a variável dependente e cada uma das variáveis independentes, cada vez mais disponibilizado de forma automática nas soluções de software de análise de dados.

Neste trabalho, seleccionaram-se estas duas medidas para avaliar os *modelos de preços hedónicos* construídos a partir da técnica de regressão linear. Descreve-se sucintamente nas próximas subsecções o que representam estas medidas.

III.5.1. Medida de avaliação da capacidade explicativa do modelo

A quantidade de variabilidade explicada (ou tida em conta) constitui uma boa medida da capacidade de ajustamento do modelo construído. O coeficiente de determinação costuma ser frequentemente utilizado como medida de avaliação em muitos trabalhos que utilizam técnicas de regressão, pelo que foi seleccionado como critério principal.

Prova-se que o coeficiente de determinação para um problema de regressão linear pode ser calculado como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \qquad \textbf{[15]}$$

Note-se que na utilização deste coeficiente deve ser alvo de alguns cuidados na sua interpretação, visto que:

- Na aplicação de uma regressão múltipla é sempre possível melhorar o coeficiente, adicionando um maior e correcto número de atributos.
- O coeficiente não é sensível à magnitude dos parâmetros *b*
- O coeficiente apresenta um erro sistemático positivo (estimativa optimista) em relação ao valor obtido quando calculado para uma dada população.

III.5.2. Medida de avaliação da relação de dependência entre *y* e *x*

A avaliação estatística da relação entre uma variável dependente e uma variável independente (numa regressão linear simples), corresponde a colocar o seguinte teste de hipótese:

$$H_0: b_1 = 0 \; vs \; H_1: b_1 \neq 0 \qquad \textbf{[12]}$$

Prova-se que a estatística de teste é dada pelo parâmetro da distribuição *t-student*, calculado tal que:

$$T_1 = \frac{\hat{b}_1}{\sqrt{\frac{SS_E}{(n-2)S_{xx}}}} \qquad \textbf{[13]}$$

Onde:

- ✓ $SS_E = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, sendo $y_i - \hat{y}_i$ designados os resíduos (e_i), que estimam o valor verdadeiro ε_i
- ✓ $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x}_i)^2$

De forma análoga, no problema de regressão linear múltipla aqui tratado, prova-se que a estatística de teste é determinada a partir de:

$$T_i = \frac{\hat{b}_i - b_i}{\hat{\sigma}\sqrt{C_{ii}}} \qquad \textbf{[14]}$$

com

- ✓ C_{ii} - o i-ésimo elemento da diagonal da matriz definida por $(\mathbf{X}'\mathbf{X})^{-1}$ (com $\mathbf{X}'$ correspondente à transposta de **X**).
- ✓ $\hat{\sigma}$ - o estimador de variância, determinado como:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} {e_i}^2}{n-a}} \qquad \textbf{[15]}$$

Este teste de hipótese permite avaliar o grau de confiança com que podemos admitir a existência de uma efectiva relação linear entre cada uma das variáveis do modelo de regressão linear múltipla construído, utilizando níveis de significância de 95% como critério mínimo de aceitação.

III.6 – Implementação em RapidMiner

A escolha da ferramenta computacional para implementar o processo de *data mining* recaiu no *software* gratuito e de código aberto *RapidMiner Community Edition* versão 5.0, disponibilizado pela empresa Rapid-I[10]. Este software implementado em JAVA possibilita a sua utilização versátil, em qualquer sistema operativo e ambiente de trabalho.

O *RapidMiner* fornece aos seus utilizadores:

i) Uma solução global para o desenho e implementação de um processo de *data mining*. Permite a definição de todas as fases incluídas na metodologia CRISP-DM, classificadas e hierarquizadas segundo as tarefas descritas: selecção de dados, integração, transformação, pré-processamento, implementação de sofisticadas técnicas de modelação, validação dos modelos e algoritmos, etc.

ii) Uma interface gráfica muito intuitiva e flexível para o desenho de um processo de *data mining*.

iii) Mais de 500 operadores que implementam as mais diversas técnicas e algoritmos. Ligação directa com a biblioteca de classes de aprendizagem automática *Weka*[11], uma das mais utilizadas na comunidade científica especializada.

iv) Acesso às mais diversas fontes de dados: Excel, Access, Oracle, IBM DB2, Microsoft SQL Server, Sybase, Ingres, mySWL, Postgres, SPSS, etc.

v) Mais de 20 métodos para visualização de dados e modelos.

vi) Repositórios de processos, dados e meta-dados.

Todos os processos de *data mining* podem ser executados na interface gráfica (modo GUI – *graphic user interface*), utilizando a linha de comandos de DOS ou acedendo via uma aplicação Java. Cada processo de *data mining* construído em *RapidMiner* é armazenado num ficheiro XML (*eXtensible Markup Language*).

[10] Mais informações em http://www.rapid-i.com
[11] Mais informações em http://www.cs.waikato.ac.nz/ml/weka/

III.6.1. Interface Gráfica do Utilizador

A organização da interface gráfica do utilizador tem disponível, de forma acessível e integradas, as diversas funcionalidades. É intuitivo construir diferentes fases e tarefas autonomamente, após definição pelo utilizador (note-se que em software de análise de dados tradicional, como por exemplo o SPSS, cada tarefa específica de um processo de análise de dados é realizada praticamente autonomamente, obrigando o investigador a efectuar muitas repetições quando pretende rever um dado processo).

Sem a necessidade de requisitos desenvolvidos em linguagem de programação para a implementação do processo, a solução fornece também um conjunto de novos algoritmos integrados em operadores. Estes novos algoritmos provêem de recentes trabalhos de investigação que abordam problemáticas específicas a certas fases do processo. Destaca-se, como exemplo, a disponibilidade de novos algoritmos de modelação como é o caso das redes neuronais ou a optimização computacional de algoritmos tradicionais, permitindo maior eficiência na utilização de volumes de dados de grande dimensão.

A interface divide-se em três perspectivas, sendo fácil e instantâneo alternar entre cada uma delas. *welcome* é o ecrã de boas-vindas. *design* é o ecrã principal onde é construído o processo de *data mining*. Por fim, o ecrã *results* possibilita a análise dos resultados, reunindo todos os outputs e possibilitando algumas ferramentas de análise gráfica. Cada perspectiva organiza-se num sistema de abas onde são disponibilizados vários tipos de vistas, seleccionáveis através do comando «VIEW → SHOW VIEW».

Centrando-nos na interface de *design*, cabe salientar que esta engloba os designados operadores. Esta designação referir-se-á aos elementos que englobam as ferramentas úteis à implementação do processo. São constituídos por uma ou mais instruções directas (escrita, concatenação, …) e implementam um ou mais algoritmos para pré-processamento, modelação e avaliação. Apresentam-se como um ícone gráfico, com um design que inclui a designação da ferramenta e o símbolo alusivo à família de operadores[12] a que pertence, bem como os conectores para os inputs / outputs de outros operadores.

[12] A família de operadores está organizada de forma próxima aos standards de *data mining*, ou seja, apresentando uma estrutura classificatória que inclui referências a fases do processo de *data mining*

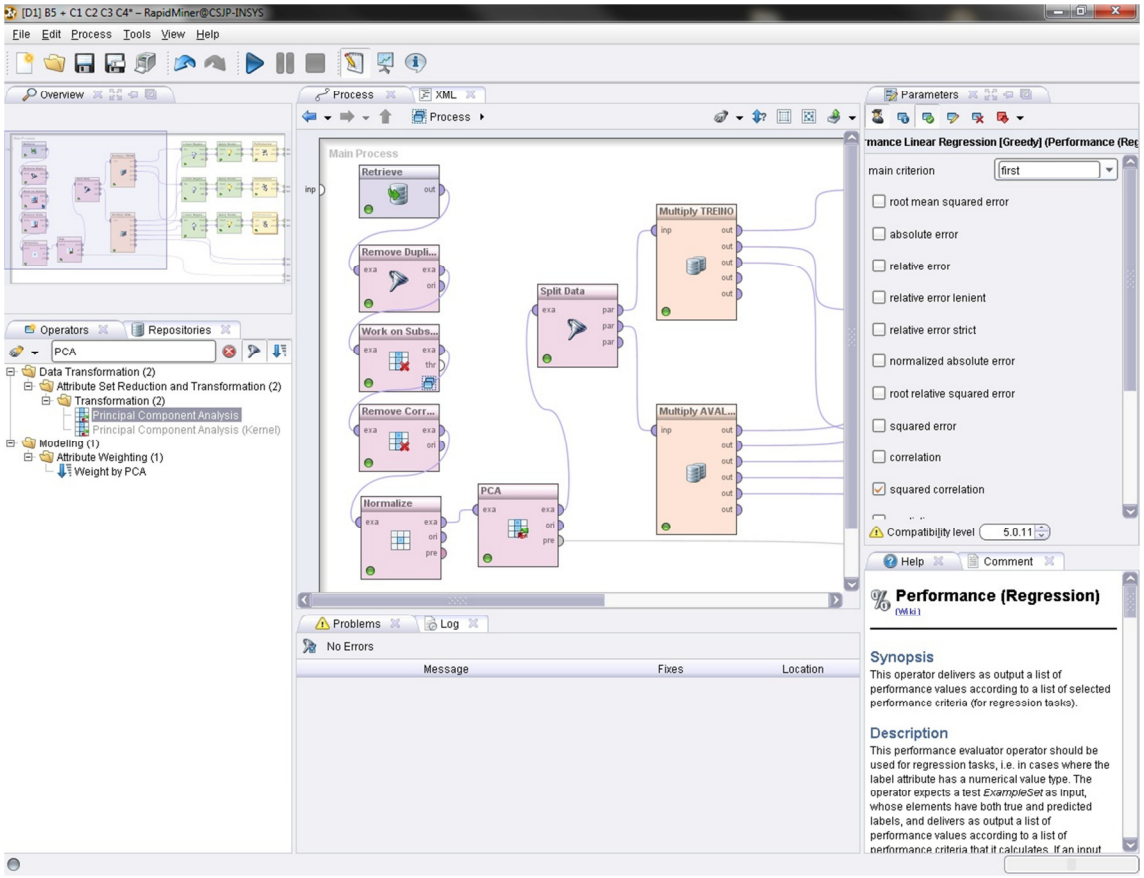

Figura 7 *Screenshot* da interface do *software RapidMiner* versão 5

Como ilustrado na Figura 7 as abas seleccionadas durante o processo de implementação aqui descrito foram sistematicamente:

- *Overview*: apresenta o esquema de implementação global de todas as ferramentas activas.
- *Operators*: lista de operadores disponíveis agrupados por categorias e subcategorias de tarefas e funções.
- *Repositories*: espaço de armazenagem das bases de dados e dos processos construídos
- *Process*: interface onde é construído o processo de forma simples, arrastando e conectando operadores
- *Tree*: interface que apresenta o processo no formato de árvore, permitindo também a construção do mesmo nesta base de visualização
- *XML*: linha de comandos da liguagem xml, utilizada pelo programa para desenhar o processo, possibilitando uma forma alternativa de construção do mesmo.

(como por exemplo, a família de operadores da classe pré – processamento), sendo também uma classificação hierárquica.

- *Problems:* aba que reúne todos os erros de concordância nos parâmetros dos operadores, identificados pelo software
- *Log*: log de todas as operações realizadas e respectiva descrição
- *Parameters*: vista que possibilita a definição de parâmetros disponíveis em cada operador
- *Help*: breve descrição de cada um dos operadores, com a descrição dos métodos algoritmos que implementa, dos parâmetros de entrada e dos outputs que fornece.
- *Comment*: permite a introdução de comentários por parte do utilizador.

III.6.2. Desenho do processo em RapidMiner

Tal como referido no tutorial[13], o desenho do processo pode ser produzido a partir da combinação de um grande número de operadores.

O *software* permite a representação do processo num sistema em árvore de operadores ou por um ambiente gráfico de fluxo de processo (*work flow*). Em ambos os casos, a estrutura do processo é ainda descrita internamente em XML, o que permite adicionalmente o desenvolvimento do processo nesta linguagem.

Apresenta ainda duas importantes funcionalidades: a possibilidade de definir pontos de interrupção do processo que permitem inspeccionar praticamente todos os resultados intermédios e a capacidade de combinar e agrupar operadores em blocos autónomos, disponíveis para processos posteriores.

A Figura 8 apresenta um processo implementado em ambiente *RapidMiner* destacando-se cinco grupos de blocos de operadores:

- ✓ bloco de pré processamento: inclui várias tarefas, englobando questões como a integração, a limpeza e a transformação de dados.
- ✓ bloco de selecção de atributos: inclui os métodos de selecção de atributos (ex: selecção por filtro)

[13] Disponível em http://sourceforge.net/projects/rapidminer/files/1.%20RapidMiner/5.0/rapidminer-5.0-manual-english_v1.0.pdf/download

- ✓ bloco de modelação: onde são inseridos os algoritmos de aprendizagem que inclui os esquemas de selecção de atributos embutidos (quando necessário).
- ✓ bloco de avaliação: onde são inseridos o esquema de validação (ex: *HoldOut*) e os operadores que permitem determinar as medidas de avaliação.

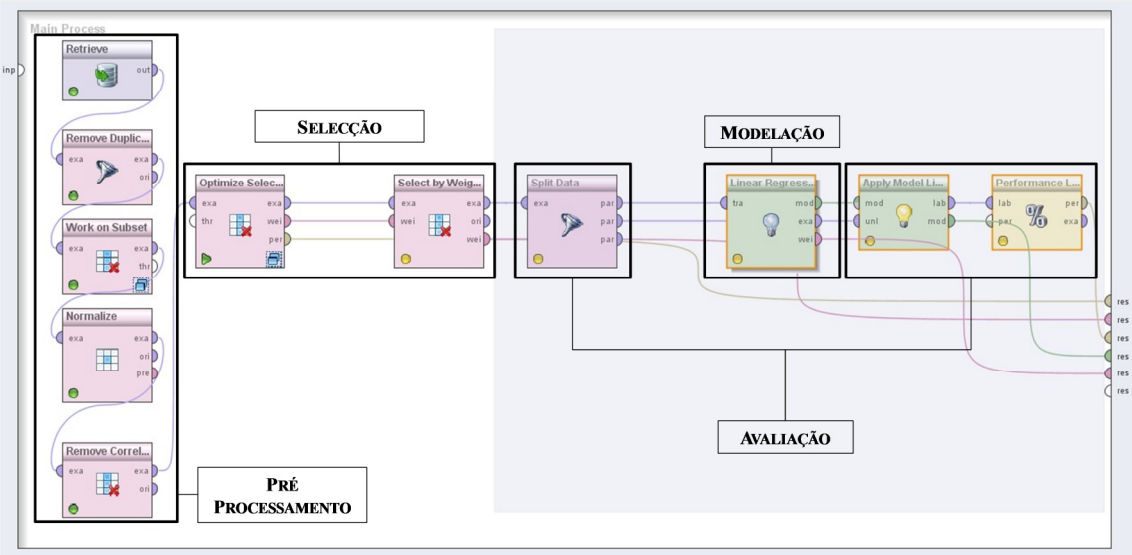

Figura 8 Exemplo da implementação de um processo em Rapid Miner. Neste processo, o pré-processamento envolve 5 tarefas, destacando-se a remoção de objectos duplicados, a normalização das variáveis e a remoção de variáveis altamente correlacionadas); a selecção é constituída por um algoritmo de busca; a aprendizagem é realizada a partir do operador de regressão linear; para a avaliação temos um esquema de validação do tipo hold-out, implementado pelo operador split data e dois operadores que permitem em conjunto determinar medidas de avaliação.

Como se descreveu anteriormente, o processo é por natureza iterativo, levando a sucessivas repetições que permitam aprimorar o resultado final.

A problemática associada à identificação de determinantes do preço da habitação aponta para a necessidade de focar o processo na selecção de variáveis que representem atributos associáveis à habitação. Estes aspectos permitiram o desenho inicial do processo, o que envolve:

1. **Fase de recolha de dados**

A recolha de dados envolve todos os procedimentos que permitem obter as bases de dados necessárias para implementar um processo que permita responder aos objectivos colocados. No presente estudo utilizaram-se bases de dados pré-existentes, disponibilizadas com elevados graus de consistência determinados pela intervenção prévia dos detentores dos dados.

2. **Fase de pré-processamento**

Nesta fase a implementação recorre às tarefas de integração, limpeza, transformação e selecção:

Recorreu-se à tarefa de integração para responder à necessidade de juntar bases de dados provenientes de diferentes fontes e ainda para formar uma base de dados global a partir de tarefas que ocorram em paralelo. O operador comum utilizado é o <JOIN>.

Para a limpeza é utilizado o operador <REMOVE DUPLICATES> de forma a evitar a existência de objectos repetidos nos dados.

- Sendo que a regressão linear impõe a inexistência de correlações entre as variáveis independentes, utiliza-se o operador <REMOVE CORRELATED ATRIBUTES> para garantir a correcta execução da fase de modelação.
- De salientar que o algoritmo de modelação inclui um mecanismo próprio embutido de eliminar as variáveis colineares (ou seja, altamente correlacionadas), que também se encontra activa.

A tarefa de transformação dos atributos engloba a implementação do operador <NORMALIZE> que efectua uma normalização das variáveis independentes reais e ordinais, recorrendo ao método *Z – transformation* (transformação da distribuição do valor dos dados numa distribuição de média zero e desvio padrão 1). Esta transformação permite eliminar as interacções associadas à escala e tipo de medida de cada uma das variáveis independentes definidas para o processo de modelação.

Para a tarefa de selecção de atributos são implementadas as 4 diferentes abordagens já descritas anteriormente, resultando num total de 8 técnicas diferentes:

i) *Redução da dimensionalidade*: utilização de operador <PCA> que implementa a análise de componentes principais; seleccionando as

componentes mais relevantes, determina-se um modelo de regressão com a utilização dos *loadings* como os novos valores associados a cada objecto

ii) *Abordagem híbrida*: redução de dimensionalidade combinada com um método de pesagem; para cada uma das componentes não descartadas da ACP, é seleccionada a variável com maior *loading* superior a 0,500. Utiliza-se o operador <WEIGHT BY PCA> para cada uma das componentes

iii) *Selecção de atributos por filtro*: utilização do operador <OPTIMIZE SELECTION> que implementa um algoritmo *greedy* tipo filtro utilizando a medida de avaliação supervisada CFS. São implementadas duas estratégias de busca:

- algoritmo *greedy* com estratégia de busca *forward*
- algoritmo *greedy* com estratégia de busca *backward*

Para calcular a medida CFS é utilizado o operador <PERFORMANCE (CFS)>.

iv) *Selecção por pesagem*:

- *utilizando ACP*: a partir da primeira componente principal obtida são seleccionadas as variáveis com um *loading* superior a 0,500. Utiliza-se o operador <WEIGHT BY PCA>
- *utilizando uma MSV linear*: como descrito na secção IV.3.1, são seleccionadas as variáveis com um coeficiente, na equação do hiperplano de separação linear, superior a 0,500. Utiliza-se o operador <WEIGHT BY SVM>

v) Selecção embutida: implementado pela activação no operador <LINEAR REGRESSION> da opção <*feature selection*>. São usados dois algoritmos:

- algoritmo de busca *greedy* com estratégia *forward* e medida de avaliação AIC
- algoritmo que implementa uma árvore de regressão M5prime.

3. Fase de modelação

Com a formulação linear do modelo de preços hedónicos, a bibliografia consultada, aponta para a utilização comum de técnicas de regressão linear multivariada. Utilizou-se um tradicional algoritmo de ajustamento, baseado no método

da determinação dos mínimos desvios quadrados, já descrito anteriormente e implementado no software pelo operador <LINEAR REGRESSION>

A combinação das fases de pré-processamento e de modelação permite implementar vários processos de *data mining*, obtendo-se 18 modelos de *preços hedónicos* para cada uma das bases de dados analisadas.

4. Fase de avaliação

Esta fase envolve o cálculo das medidas de avaliação descritas na secção IV.5 e obtidas a partir dos operadores <APPLY MODEL> (para as estatísticas de teste do teste de hipótese da efectiva relação linear de cada um dos atributos com a variável dependente) e <PERFORMANCE LINEAR REGRESSION> (para o coeficiente de determinação R^2). Estas medidas são estimadas recorrendo a dois possíveis esquemas de validação, utilizando como critério o número global de dados e o necessário menor custo computacional. A escolha será realizada entre os dois métodos já referidos, sendo o método *hold-out* implementado com o operador <SPLIT VALIDATION> ou o método de validação cruzada *k-fold* implementado com o operador <X VALIDATION> para um k=10.

IV. O MERCADO DA HABITAÇÃO À ESCALA NACIONAL

Os fenómenos socioeconómicos exercem impactos significativos nos padrões territoriais da habitação. É expectável que estes padrões sejam detectáveis no mercado global da habitação, onde são traduzidos nos mecanismos de formação de preços.

O enquadramento teórico deste trabalho permitiu rever as principais dinâmicas socioeconómicas com expectáveis consequências nos actuais padrões da habitação [capítulo II e subsecção II.1]. A construção de um *modelo de preços hedónicos* que incorpore, de alguma forma, a dimensão espacial permite-nos identificar a possível relação causal entre os fenómenos e os atributos, determinantes do preço da habitação à macro escala.

Identificar esta relação entre os fenómenos socioeconómicos e os atributos identificados, associada à reflexão sobre os resultados destes modelos à macro – escala, permite-nos uma primeira validação da implementação global do processo e da efectiva capacidade da conjugação das ferramentas econométricas com um processo de *data mining*.

Concentrando as atenções exclusivamente nos atributos físicos da habitação, o INE fornece um conjunto largo de indicadores com uma desagregação espacial máxima ao nível do município. Esta é uma unidade espacial de carácter administrativo suficientemente grande para abarcar fenómenos socioeconómicos de carácter nacional mas que ao mesmo tempo fornece um número razoável de graus de liberdade para desenvolver o caso de estudo a partir da implementação de um processo de *data mining*.

O objectivo deste caso de estudo centra-se na identificação de quais os indicadores das características da habitação, ao nível municipal, que são efectivamente determinantes na formação do preço médio de transacção nos prédios nos municípios de Portugal continental.

IV.1 – Dados disponíveis

A informação censitária constitui uma das principais fontes de dados produzidos pelo INE. A informação é disponibilizada agregada territorialmente, embora permitindo consultas até ao nível de freguesia.

Foram compiladas 28 variáveis da categoria *habitação e construção* da base de dados online[14]. Destas variáveis 30% referem-se a aspectos que caracterizam genericamente a vizinhança das habitações[15], envolvendo questões como a densidade de edificado, o número total de habitações e questões relativas à ocupação dos edifícios. As restantes 70% definem aspectos estruturais dominantes.

Das variáveis anteriores, uma refere-se ao indicador de preço da habitação. O único disponível reporta o valor médio da transacção de prédios[16] em cada município, sendo escolhido com a referência ao ano de 2001.

Seleccionadas para o menor nível de detalhe comum (nível municipal de Portugal Continental, *N*=278), referem-se aos dados compilados dos Censos de 2001 e de 1991 (sendo que a referência a dados de 1991 é exclusiva à variável *variação do número total de habitações no período de 1991 a 2001*).

[14] www.ine.pt
[15] O INE utiliza o conceito de alojamento como uma referência lata ao local que se destina a habitação, o que permite tratar a expressão como equivalente ao sentido dado à designação habitação neste trabalho.
[16] Segundo o INE, o conceito de prédio refere-se à parte delimitada do solo juridicamente autónoma, abrangendo as águas, plantações, edifícios e construções de qualquer natureza nela incorporados ou assentes com carácter de permanência; é ainda considerado prédio cada fracção autónoma no regime de propriedade horizontal.

IV.2 – Constrangimentos associados às variáveis recolhidas

IV.2.1. Limitações associadas a variáveis independentes

Os dados disponibilizados pelo INE referem-se a valores absolutos de cada um dos atributos seleccionados. A utilização directa deste tipo de valores não se revela adequada neste estudo uma vez que as unidades territoriais de análise (os municípios) não têm delimitações homogéneas. Consequentemente, é expectável que os valores absolutos reflictam a dimensão territorial, a qual constitui um factor dominante. Neste caso, os valores absolutos de qualquer variável são directamente correlacionados com a dimensão espacial que reportam e inviabilizam a caracterização da realidade estrutural dos atributos da habitação – factores menos relevantes.

O indicador de preço utilizado refere-se a um valor médio, para uma dada delimitação territorial. Por uma questão de consistência, a possível análise comparativa municipal exige a utilização de variáveis, independentes, ponderadas pelo seu factor quantitativo, dado pela dimensão territorial.

A excepção à regra anterior refere-se à utilização de uma variável que sirva de indicador dimensional (não o expurgando totalmente do modelo). O número total de habitações é um indicador que nos permite representar grande parte do efeito quantitativo, dado pela dimensão territorial, e ao mesmo tempo, avaliar de que forma o volume de habitações é um factor determinante na formação do valor médio da transacção dos prédios em cada município.

IV.2.2. Limitações associadas a variável dependentes - indicador de preço da habitação

A utilização, como variável dependente, do valor médio da transacção de prédios envolve várias restrições.

O conceito de prédio engloba, para além das habitações, outros bens imobiliários: por exemplo os edifícios não residenciais e as parcelas de solo desocupadas.

A ausência de outro indicador alternativo, disponibilizado pelo INE obrigou à utilização desta variável, embora com a limitação óbvia associada. Apontam-se os aspectos que poderão influenciar a interpretação dos modelos obtidos:

- A maior ou menor dinâmica do volume de transacções na unidade territorial é outro problema na utilização deste indicador. Um menor volume de transacções é sensível a dados com grande variabilidade. Por exemplo, um município em que se transaccione um único prédio, localizado no centro de uma área urbana e de parcas dimensões pode apresentar um valor bastante elevado embora comparável, na mesma ordem de grandeza, a um prédio que abarque exclusivamente funções agrícolas, de dimensão assinalável, numa outra unidade territorial.
- As características funcionais da unidade territorial exercem uma influência importante nos valores médios das transacções. Como alertam vários autores, o mercado imobiliário engloba vários submercados com características próprias. É expectável que funções residenciais e não residenciais, por exemplo, representem realidades muito distintas. Dessa forma, os valores médios de prédios transaccionados num dado município poderão apresentar uma relação directa com a estrutura espacial das funções existentes.

IV.3. – Implementação do processo de data mining

O processo implementado seguiu o desenho proposto na secção III.6.2. Para este primeiro caso de estudo, as adaptações consideradas para as diferentes fases englobam:

Objectivo: identificar os aspectos que influenciam a formação do preço da habitação ao nível municipal, partindo da identificação dos atributos determinantes do valor médio dos prédios transaccionados.

Pré-processamento: para além das tarefas descritas anteriormente, o pré processamento dos dados envolveu a transformação das variáveis absolutas, para uma ponderação relativa dos valores absolutos globais – esta tarefa foi realizada antes da integração da base de dados no RapidMiner, tendo-se utilizado o software MS Excell.

Modelação: não existiu nenhuma diferença em relação ao processo desenhado para a fase de modelação. Aplicou-se o operador de regressão linear.

Avaliação: dado o número reduzido de casos, optou-se por um esquema de validação que implementa um esquema *k-fold*, com $k = 10$ correspondendo ao número de partições em que é subdividido o conjunto de dados. Com este esquema, a medida de avaliação dada pelo coeficiente de determinação R^2 é obtida pela média aritmética dos coeficientes de determinação em cada iteração. O processo iterativo calcula o R^2 a partir de um modelo obtido no conjunto de treino constituído por 9 partições e, utilizando a partição restante, obtém a medida de avaliação parcial – o processo é repetido para todas as combinações possíveis de obter com as 10 partições iniciais.

IV.3.1. Descrição da base de dados

Na Tabela 1 apresentam-se os dados iniciais utilizados, bem como as principais características que lhe estão associadas.

Grande parte das variáveis recolhidas possui unidades de medida semelhantes, associadas a duas dimensões intrínsecas: uma quantitativa e outra categórica. A dimensão quantitativa refere-se à quantidade de habitação e a categórica está associada ao tipo de conceito a que se refere a quantificação. Estes conceitos, como são exemplos a propriedade, o uso da habitação e o número de pisos desdobram-se em categorias como é exemplo, para a propriedade, se esta é atribuída ao próprio, a uma entidade pública ou a outra entidade privada.

As variáveis recolhidas englobam duas das dimensões de análise definidas em III.4.3 / equação [3]: os atributos estruturais, assinalados com um (E) na última coluna da Tabela 1e os atributos de vizinhança, assinalados com um (V).

Sendo variáveis agregadas podem descrever vários conceitos. A categorização estrutural ou vizinhança não é estanque.

Não foi incorporada uma variável relativa à localização, uma vez que este é um conceito intrínseco aos dados, que se referem a delimitações municipais.

Tabela 1 Descrição dos atributos da base de dados

CODIGO	DESCRIÇÃO	PAPEL NO MODELO	TIPO VAR	UNID MED	ESTATÍSTICAS			TIPO ATR
					MÉDIA E DESVIO PADRÃO	MÍN	MÁX	
ID		id	nominal	-	-	-	-	-
CONC_NAME	Designação do município	grupo	nominal	-	-	-	-	-
V90_Valor	Valor médio dos prédios transaccionados	variável dependente	real	€	δ = 34509,0 +/- 24214,5	4001,0	181777,0	-
V01_Alojamentos	Número total de habitações	variável independente	real	Nº	δ = 12268,2 +/- 20550,4	706,0	221868,0	V
V02_VarAloj9101	Variação do número total de habitações no período de 1991 a 2001	variável independente	real	%	δ = 16,3 +/- 12,4	-6	66	V
V03_Uso1ªHab	Percentagem de habitações com utilização como primeira habitação	variável independente	real	%	δ = 65,1 +/- 10,7	37	89	V
V04_UsoSazonal	Percentagem de habitações com utilização sazonal	variável independente	real	%	δ = 24,1 +/- 11,2	5	54	V
V05_UsoVago	Percentagem de habitações sem utilização (vagas)	variável independente	real	%	δ = 10,1 +/- 2,9	3	21	V
V06_AlojSublotado	Percentagem de habitações sublotadas	variável independente	real	%	δ = 62,8 +/- 8,8	42	84	E

V07_AlojSuperlotado	Percentagem de habitações superlotadas	variável independente	real	%	δ = 13,5 +/- 4,6	5	28	E
V08_PropOcupante	Percentagem de habitações propriedade do próprio ocupante	variável independente	real	%	δ = 83,4 +/- 9,3	48	99	E
V09_PropEntPrivada	Percentagem de habitações propriedade de uma entidade privada não ocupante	variável independente	real	%	δ = 11,6 +/- 6,6	1	34	E
V10_PropEntPublica	Percentagem de habitações propriedade de uma entidade pública não ocupante	variável independente	real	%	δ = 1,9 +/- 2,1	0	16	E
V11_AlojEdif	Quociente do número de alojamentos pelo número total de edifícios existentes	variável independente	real	Nº	δ = 1,3 +/- 0,6	1	6	V
V12_EdifKm	Densidade de edifícios no município	variável independente	real	Nº	δ = 144,7 +/- 409,3	5	3463	V
V13_Edif1piso	Percentagem de edifícios com apenas 1 piso	variável independente	real	%	δ = 43,6 +/- 22,2	6	95	E
V14_Edif2pisos	Percentagem de edifícios com 2 pisos	variável independente	real	%	δ = 44,8 +/- 18,5	5	79	E
V15_Edif3pisos	Percentagem de edifícios com 3 pisos	variável independente	real	%	δ = 8,3 +/- 6,9	0	48	E
V16_Edif6oumaispisos	Percentagem de edifícios com 6 ou mais pisos	variável independente	real	%	δ = 1,5 +/- 3,3	0	27	E
V17_EdifFuncNaoRes	Percentagem de edifícios com funções não residenciais	variável independente	real	%	δ = 11,5 +/- 7,9	0	50	V
V18_ConstrAntes60	Percentagem de edifícios com data de construção anterior a 1960	variável independente	real	%	δ = 22,7 +/- 6,4	6	47	E
V19_Constr6190	Percentagem de edifícios com data de construção referente ao período de 1961 a 1990	variável independente	real	%	δ = 49,0 +/- 8,4	24	73	E
V20_Constr9101	Percentagem de edifícios com data de construção referente ao período de 1991 a 2001	variável independente	real	%	δ = 19,0 +/- 4,1	7	35	E
V21_IdadeMedia	Idade média dos edifícios	variável independente	real	Nº	δ = 35,3 +/- 6,0	19	57	E
V22_NumDivisoes	Número médio de divisões das habitações	variável independente	real	Nº	δ = 4,7 +/- 0,3	3,9	5,6	E
V23_EdifDegradados	Percentagem de edifícios classificados como degradados	variável independente	real	%	δ = 3,0 +/- 1,7	0,1	11,5	E
V24_EdifAqueCentral	Percentagem de edifícios com aquecimento central	variável independente	real	%	δ = 5,2 +/- 4,1	0,1	19,3	E
V25_EdifRSU	Percentagem de edifícios servidos por sistema de recolha de resíduos sólidos urbanos	variável independente	real	%	δ = 90,3 +/- 8,7	50,5	99,7	E
V26_AlojInfraBasica	Percentagem de habitações sem acesso a pelo menos uma infraestrutura básica (água, saneamento, electricidade)	variável independente	real	%	δ = 13,0 +/- 6,1	2,2	38,6	E
V27_EdifAcessivel	Percentagem de edifícios classificados como acessíveis a pessoas de mobilidade reduzida	variável independente	real	%	δ = 66,3 +/- 15,6	9,9	99,1	E

IV.4 – Resultados

IV.4.1. Análise global

O processo inicial de limpeza envolveu a aplicação dos operadores <REMOVE DUPLICATES> e <REMOVE CORRELATED ATRIBUTES>. A remoção de variáveis altamente correlacionadas resultou na eliminação das variáveis V04_USO SAZONAL, V09_PROPRIEDADE ENTIDADE PRIVADA e V11_QUOCIENTE ALOJAMENTOS EDIFÍCIOS.

Para além da capacidade de recolher os atributos correctos por parte do investigador, a tarefa de selecção, no processo de *data mining*, assume grande relevância: permite reduzir ainda mais o conjunto de atributos inicial e proporcionar a capacidade de o algoritmo de aprendizagem identificar aqueles que são efectivamente determinantes do valor médio dos prédios nos municípios portugueses.

Para determinar a mais-valia da utilização de ferramentas de selecção, implementou-se processos com diferentes algoritmos de selecção e efectuando uma análise comparativa com um modelo de base. Este modelo de base é obtido do conjunto de atributos da habitação recolhidos. Com as variáveis inicialmente removidas envolve um total de 24 atributos.

Os restantes 8 modelos implementam diferentes algoritmos, integrados nas 4 grandes abordagens atrás descritas: selecção embutida, selecção filtro, selecção por pesagem e redução da dimensionalidade.

A Tabela 2 indica-nos os resultados globais dos diferentes modelos obtidos.

Tabela 2 Resultados globais dos diferentes *modelos de preços hedónicos* construídos

	SEM SELECCÇÃO	COM SELECÇÃO							
		Abordagem embutida		Redução da dimensionalidade		Abordagem Filtro		Técnicas de pesagem	
		Árvore de regressão [M5 Prime]	Algoritmo Greedy [estratégia forward e critério AIC]	Com novas variáveis	Com variáveis representativas dos conceitos	Heurística FSS [estratégia backward e critério CFS]	Heurística FSS [estratégia forward e critério CFS]	PCA *Weighting*	SVM *Weighting*
Modelos	**M0**	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
Variáveis seleccionadas	**24**	14	10	5	5	**7**	7	12	**12**
Variáveis significantes	**14**	11	10	3	4	**7**	7	9	**9**
Capacidade explicativa	**0,742**	0,730	0,746	0,775	0,471	**0,762**	0,764	0,742	**0,770**
Variação no número de variáveis	-	-41,7%	-58,3%	-79,2%	-79,2%	**-70,8%**	-70,8%	-50,0%	**-50,0%**
Variação da capacidade explicativa	-	-1,6%	0,5%	4,4%	-36,5%	**2,7%**	3,0%	0,0%	**3,8%**

Para a avaliação dos modelos apresentam-se indicadores relativos ao número de atributos seleccionados por cada um dos algoritmos de selecção implementados (após as restantes fases de pré – processamento).

A efectiva relação linear de cada um dos atributos com o valor médio dos prédios transaccionados, nos vários modelos finais, é determinada pelo teste de significância (estatística T), assumindo-se como significantes para níveis de confiança mínimos de 95%.

O coeficiente de determinação R^2 oferece-nos uma avaliação da capacidade explicativa do modelo, que, neste caso, varia entre os 74% do modelo de base e os 76% do modelo obtidos a partir da transformação dos atributos iniciais num conjunto de 5 novas variáveis, obtidas por combinação linear, através da ACP.

Na generalidade, verifica-se que a utilização das técnicas de selecção permite diminuir o número de atributos que fazem parte do modelo de preços hedónicos sem perdas significativas na capacidade explicativa do valor dos prédios. A maioria das técnicas consegue, efectivamente, melhorar a capacidade explicativa em relação ao modelo inicial, o que demonstra a existência de atributos irrelevantes no conjunto de variáveis compiladas – a este facto não é alheio o elevado número de atributos (10) não significantes para os níveis de confiança de 95%, no modelo de base.

IV.4.2. Análise da capacidade explicativa e número de variáveis

Dos modelos construídos, destaca-se a utilização da técnica de redução de dimensionalidade ACP, utilizando as novas variáveis (modelo M3) obtidas por transformação linear. Com uma capacidade explicativa do valor de transacção médio dos prédios ao nível municipal de 77%, o resultado global, em termos comparativos, aponta para um acréscimo da capacidade explicativa de 4,4% ao mesmo tempo que se pode atribuir uma redução de 80% no número de variáveis necessárias para descrever o problema – ambos os indicadores são relativos ao modelo de base (modelo M0).

Tabela 3 Análise de Componentes Principais para o conjunto de 24 variáveis inicial

Componente	Valores Próprios	Variância Explicada	Variância cumulativa
PC 1	2,363	0,233	0,233
PC 2	2,174	0,197	0,430
PC 3	1,692	0,119	0,549
PC 4	1,446	0,087	0,636
PC 5	1,016	0,043	0,679
(…)	(…)	(…)	(…)

Variáveis	C1	C2	C3	C4	C5
V01_Alojamentos	0,520	0,515			
V02_VarAloj9101		0,578			
V03_Uso1ªHab					
V05_UsoVago					
V06_AlojSublotado	-0,664				
V07_AlojSuperlotado					
V08_PropOcupante	-0,737				
V10_PropEntPublica	0,579				
V12_EdifKm	0,532		-0,540		
V13_Edif1piso	0,607	-0,546			
V14_Edif2pisos	-0,749				
V15_Edif3pisos			-0,530		
V16_Edif6oumaispisos	0,539				
V17_EdifFuncNaoRes					0,639
V18_ConstrAntes60		-0,672			
V19_Constr6190		0,746			
V20_Constr9101			0,684		
V21_IdadeMedia		-0,750			
V22_NumDivisoes	-0,818				
V23_EdifDegradados				0,557	
V24_EdifAqueCentral	-0,584				
V25_EdifRSU				-0,586	
V26_AlojInfraBasica				0,581	
V27_EdifAcessivel					

Como se pode verificar na Tabela 3, a extracção das cinco primeiras componentes principais, com valores próprios superiores a 1, explicam cerca de 68% da variância contida nos dados iniciais. Para a percepção do significado do modelo é necessário descrever empiricamente os conceitos subjacentes a cada componente. A matriz de *loadings* (Tabela 3) é essencial nesta análise.

Eliminando os valores inferiores a 0,500 – o que permite destacar as variáveis mais importantes associadas a cada componente – procura-se, através de uma análise empírica, atribuir um significado aos novos conceitos, latentes nos dados iniciais:

- **Componente 1**: Volume de espaço construído
- **Componente 2**: Idade do edificado
- **Componente 3**: Dinâmica quantitativa da habitação
- **Componente 4**: Qualidade da habitação
- **Componente 5**: Funções dominantes presentes nos edifícios

Porém, este modelo deve ser considerado com algum cuidado visto que é comum existirem algumas limitações, destacando-se:

- ✓ A interpretação empírica das novas variáveis nem sempre é fácil para o investigador. Exige um conhecimento teórico aprofundado sobre a temática, que permita relacionar a combinação de atributos iniciais num determinado conceito mais geral. No presente caso, as variáveis são todas perfeitamente identificáveis e associadas a conceitos globais facilmente identificados.
- ✓ A construção de novas variáveis, por si só, não permite seleccionar as variáveis iniciais. Assim, a técnica não permite determinar parte do problema da identificação efectiva das variáveis iniciais que são determinantes do preço, visto que todas contribuem para a construção do modelo de ACP.
- ✓ A ACP é um modelo dos dados não supervisionado. Desta forma, a recolha e selecção, a *priori*, das variáveis iniciais influencia o tipo de modelo obtido e que pode resultar num modelo reduzido pouco perceptível para a temática em estudo.

Como vemos no modelo de ACP obtido, encontramos variáveis iniciais como a referente aos edifícios classificados como acessíveis a pessoas de mobilidade reduzida, ao uso da habitação como primeira habitação ou às habitações vagas, que não têm uma

associação clara a nenhuma das componentes. Mesmo nas restantes componentes e, tirando algumas das variáveis da primeira componente não têm uma relação extremamente forte com a respectiva componente (na maioria, a percentagem de variância de cada variável que contribui para a definição de uma dada componente gira em torno dos 55%). Estes aspectos introduzem maior grau de incerteza na interpretação do modelo de regressão uma vez que não é possível afirmar de forma unívoca que a utilização de outro conjunto de variáveis inicial (recolhidas pelo investigador) não introduz diferenças na capacidade explicativa e número de variáveis.

Este facto alerta-nos ainda para a importância de técnicas complementares, como as técnicas de rotação ortogonal, frequentemente utilizadas na ACP de forma a facilitar a interpretação das componentes e a aceitabilidade da ACP para a descrição da problemática em estudo. No entanto, este tipo de técnicas não se encontra disponível através de um operador próprio no *software* utilizado.

Nos restantes modelos construídos, destaca-se o modelo obtido após da aplicação de um esquema de pesagem com máquinas de suporte vectorial (SVM) (modelo **M8**) para a selecção de atributos relevantes. Este modelo apresenta uma capacidade explicativa de 75% com a vantagem de proporcionar a utilização de variáveis originais (não transformadas).

Os modelos obtidos após a implementação de um método de selecção, de tipo filtro, com algoritmo de busca *greedy* e medida de avaliação CFS (modelos **M5** e **M6**), obtêm a melhor eficiência de selecção (menor número de atributos). Com estes mecanismos de selecção obtém-se uma redução em torno de 71% das variáveis iniciais, sendo que os *modelos de preços hedónicos* resultantes apresentam melhorias na capacidade explicativa de 3% em relação ao modelo base. Este é um valor (76%) semelhante ao obtido com o modelo **M8** (75%).

IV.4.3. Análise das variáveis seleccionadas

A Tabela 4 apresenta, para cada uma das variáveis seleccionadas e cada um dos modelos construídos, os respectivos *coeficientes estandardizados.* Os coeficientes

revelam-nos a variação, medida em desvio padrão, da variável dependente quando existe a variação de uma unidade do desvio padrão da variável independente respectiva.

A análise destes coeficientes permite-nos avaliar as variáveis seleccionadas sistematicamente pelos diferentes algoritmos de selecção, o que se considerou um indicador aceitável do nível de importância de cada atributo, para a problemática de estudo. Esta é uma análise comparativa *ad-hoc* que complementa a análise quantitativa proporcionada pela avaliação dos valores dos coeficientes (estandardizados[17]) e a avaliação da importância relativa de cada atributo em cada um dos modelos construídos.

Como se pode verificar na Tabela 4, não existem diferenças assinaláveis na hierarquia dos modelos mais eficientes atrás mencionados. Sistematicamente, é a mesma variável independente – quando seleccionada previamente – que induz a maior variação da variável dependente. De forma complementar, é possível quantificar o número de vezes que cada variável incorpora cada um dos modelos A última coluna da tabela refere-se ao número global de variáveis sistematicamente incluídas e significantes, dando uma pista da sua real relevância para o tema em estudo.

Tabela 4 Quadro síntese dos coeficientes estandardizados associados a cada uma das variáveis, para cada um dos *modelos de preços hedónicos*.

	SEM SELECÇÃO **MODELO BASE**	COM SELECÇÃO							
		Abordagem embutida		Redução da dimensionalidade		Abordagem Filtro		Técnicas de pesagem	
		Árvore de regressão [M5 Prime]	Algoritmo Greedy [estratégia forward e critério AIC]	Com novas variáveis	Com variáveis representativas dos conceitos	Heurística FSS [estratégia backward e critério CFS]	Heurística FSS [estratégia forward e critério CFS]	PCA *Weighting*	SVM *Weighting*
Modelos	**M0**	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
Capacidade explicativa	**0,742**	0,730	0,746	**0,775**	0,471	0,762	0,764	0,742	**0,770**
V01_Alojamentos	**0,107**	0,135	0,132	-		0,112	0,114	0,118	**0,112**
V02_VarAloj9101	**0,086**	0,080	0,098	-		0,133	0,126		**0,100**
V03_Uso1ªHab	**0,067**			-					
V04_UsoSazonal				-					0,015

[17] Uma vez que as variáveis independentes são normalizadas, a informação associada aos coeficientes não estandardizados não é significativamente diferente dos coeficientes estandardizados para esta análise.

V05_UsoVago	-0,027			-					
V06_AlojSublotado	**-0,421**	-0,521	-0,543	-		-0,180	-0,186	-1,024	**-0,681**
V07_AlojSuperlotado	**-0,295**	-0,347	-0,387	-				-0,685	**-0,414**
V08_PropOcupante		-0,218	-0,211	-		-0,194		-0,204	
V09_PropEntPrivada	**0,187**			-			0,207		**0,196**
V10_PropEntPublica	**0,073**			-				0,015	
V11_AlojEdif	**0,471**	0,385	0,322	-		0,295	0,284		
V12_EdifKm	0,002			-				-0,044	
V13_Edif1piso	**0,825**	0,474	0,140	-				-0,018	**0,087**
V14_Edif2pisos	**0,533**	0,249		-		-0,208	-0,206	-0,127	-0,047
V15_Edif3pisos	**0,278**	0,149		-					
V16_Edif6oumaispisos				-				0,333	**0,289**
V17_EdifFuncNaoRes	0,028			-	0,067				
V18_ConstrAntes60	-0,051		-0,068	-					
V19_Constr6190	-0,043	0,058		-					
V20_Constr9101	0,010	0,053		-	-0,106				
V21_IdadeMedia	**-0,064**			-	-0,297				
V22_NumDivisoes	-0,043			-	-0,688			0,101	0,054
V23_EdifDegradados	0,009			-					
V24_EdifAqueCentral	0,025			-				0,064	
V25_EdifRSU	0,047	0,047		-	0,207				
V26_AlojInfraBasica	**-0,094**	-0,102	-0,128	-		-0,171	-0,170		**-0,144**
V27_EdifAcessivel	**0,068**	0,055	0,058	-				0,123	**0,063**

Os resultados apresentados nesta tabela permitem classificar as variáveis relevantes em três níveis:

- **NÍVEL 1 – Extremamente Relevantes**
  - V01 e V06: [*volume total de habitações* e percentagem de alojamentos sublotados]
- **NÍVEL 2 – Muito Relevantes**
  - V02 e V26: [*taxa de variação do número de habitações entre* 91 *e* 2001 *e a percentagem de habitações não cobertas por pelo menos uma infraestrutura básica*]
- **NÍVEL 3 – Relevantes**
  - V07, V11 e V14: [*percentagem de alojamentos superlotados, densidade de habitações pelo total de edifícios* e *percentagem de edifícios de 2 pisos*]

As variáveis V01 e V06 podem ser consideradas extremamente relevantes sendo as que mais influenciam o valor médio dos prédios transaccionados. Como observado na última coluna da Tabela 4, estas duas variáveis formam parte de 7 dos 8 modelos construídos. No caso do número total de habitações [V01], o aumento de uma unidade de desvio padrão induz uma variação, no valor médio dos prédios transaccionados, de 12% do seu desvio padrão. Por seu lado, a percentagem de alojamentos sublotados [V01] apresenta uma variabilidade do coeficiente expressiva de modelo para modelo. No entanto é, sistematicamente, a segunda variável mais importante na ordem dada estabelecida pelo módulo dos coeficientes de cada modelo individualmente. O impacto do aumento de um desvio padrão desta variável é, em média, associado a um decréscimo médio no valor médio dos prédios transaccionados, de 51% do seu desvio padrão.

No seguinte nível, podem ser consideradas como muito relevantes as variáveis V02 e V26 seleccionadas em 6 dos 8 modelos considerados. O aumento de um desvio padrão na taxa de variação do número de alojamentos de um dado município [V02] representa um acréscimo de 10% do desvio padrão do valor médio das transacções. Já a inexistência de pelo menos uma infraestrutura [V26] é responsável pelo decréscimo de 14%. O valor dos coeficientes e a sua importância relativa na ordem de cada modelo não variam de forma expressiva – ocupando um lugar modesto nos impactos expectáveis no valor médio dos prédios transaccionados.

Finalmente como relevantes aparecem as variáveis V07, V11 e V14, seleccionadas em 5 dos 8 modelos resultantes. Apesar de aparecer apenas neste último nível, a variável relativa à percentagem de alojamentos superlotados [V07] apresenta o maior impacto no valor médio dos prédios transaccionados: um decréscimo de 43%, do desvio padrão, no valor médio dos prédios transaccionados.

IV.4.4. Análise dos níveis de significância

A Tabela 5 apresenta, para cada uma das variáveis seleccionadas em cada um dos modelos construídos, os respectivos *p-values* para o teste de hipótese, da efectiva relação linear de cada uma das variáveis independentes e da variável dependente [que, como referido, é determinado automaticamente no operador <APPLY MODEL>.

Como se pode identificar, o modelo inicial é aquele que apresenta um maior número de variáveis não significantes, o que aponta para a maior dificuldade em identificar os atributos determinantes do valor médio de transacção quando aplicado um modelo linear (10 variáveis não significantes).

As técnicas de pesagem e a árvore de regressão embutida, também revelam alguma fragilidade, uma vez que dos conjuntos de atributos seleccionados resultam modelos em que nem todas as variáveis são realmente significantes.

Tabela 5 Quadro síntese dos *p-values* associados a cada uma das variáveis para cada um dos modelos construídos.

	SEM SELECÇÃO **MODELO BASE**	COM SELECÇÃO							
		Abordagem embutida		Redução da dimensionalidade		Abordagem Filtro		Técnicas de pesagem	
		Árvore de regressão [M5 Prime]	Algoritmo Greedy [estratégia forward e critério AIC]	Com novas variáveis	Com variáveis representativas dos conceitos	Heurística FSS [estratégia backward e critério CFS]	Heurística FSS [estratégia forward e critério CFS]	PCA *Weighting*	SVM *Weighting*
Modelos	**M0**	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
Capacidade explicativa	**0,742**	0,730	0,746	**0,775**	0,471	0,762	0,764	0,742	**0,770**
V01_Alojamentos	**0,000**	0,000	0,000			0,000	0,000	0,000	**0,000**
V02_VarAloj9101	**0,002**	0,004	0,000			0,000	0,000		**0,000**

V03_Uso1ªHab	**0,019**								
V04_UsoSazonal									0,604
V05_UsoVago	0,343								
V06_AlojSublotado	**0,000**	0,000	0,000			0,000	0,000	0,000	**0,000**
V07_AlojSuperlotado	**0,000**	0,000	0,000					0,000	**0,000**
V08_PropOcupante		0,000	0,000			0,000		0,000	
V09_PropEntPrivada	**0,000**						0,000		**0,000**
V10_PropEntPublica	**0,010**							0,603	
V11_AlojEdif	**0,000**	0,000	0,000			0,000	0,000		
V12_EdifKm	0,938							0,170	
V13_Edif1piso	**0,000**	0,000	0,000					0,537	**0,002**
V14_Edif2pisos	**0,000**	0,000				0,000	0,000	0,000	0,108
V15_Edif3pisos	**0,000**	0,000							
V16_Edif6oumaispisos								0,000	**0,000**
V17_EdifFuncNaoRes	0,460				0,156				
V18_ConstrAntes60	0,083		0,015						
V19_Constr6190	0,158	0,043							
V20_Constr9101	0,722	0,064			0,017				
V21_IdadeMedia	**0,027**				0,000				
V22_NumDivisoes	0,153				0,000			0,001	0,060
V23_EdifDegradados	0,743								
V24_EdifAqueCentral	0,378							0,033	
V25_EdifRSU	0,112	0,104			0,000				
V26_AlojInfraBasica	**0,001**	0,000	0,000			0,000	0,000		**0,000**
V27_EdifAcessivel	**0,018**	0,056	0,040					0,000	**0,027**

O facto de os conjuntos de atributos resultarem em modelos com variáveis não significantes pode ficar a dever-se a dois constrangimentos:

i) Conjunto de dados iniciais com variáveis irrelevantes para o problema em estudo, que resulta também em maiores desvios entre os valores reais da variável dependente e os valores preditos pelo ajustamento determinado;

ii) Existência de outro tipo de formulação (não linear) na relação entre as variáveis independentes e a variável dependente, não testada nos ajustamentos construídos.

V. O MERCADO DA HABITAÇÃO À ESCALA LOCAL

Replicar o estudo anterior para um maior detalhe, que nos permita estudar a habitação como unidade de análise individual, é um problema de investigação que assume maior complexidade.. Nesse sentido, a realidade territorial dada pelos municípios de Aveiro e Ílhavo, alvo de diversos trabalhos ao longo do meu percurso académico, são escolhas óbvias para a constituição de um caso de estudo.

Para este trabalho, à micro escala, foi necessário aceder a informação sobre habitações transaccionadas no mercado. Nesse sentido, a colaboração da empresa Janela Digital, responsável pelo portal Casa Sapo, foi essencial ao fornecer dados relativos a um grande número de habitações transaccionadas nos dois municípios referidos.

Serviços de informação geográfica como o Sapo Mapas, albergam uma quantidade elevada de informação relativa ao território. Como vimos, a habitação está intimamente ligada ao local onde se encontra e, dessa forma, às características da sua envolvente. A utilização complementar da informação georreferenciada disponibilizada pelo LabSapo – Laboratório Sapo da Universidade de Aveiro – constituiu a oportunidade de alargar a pesquisa dos atributos determinantes do preço da habitação, acrescentando possíveis relações com a oferta de equipamentos e serviços suas proximidades.

A disponibilização dos dados, armazenados nos últimos 10 anos, das habitações publicitadas no portal Casa Sapo, procurou-se testar as ferramentas utilizadas no caso de estudo anterior para níveis de complexidade muito mais elevados – com um maior número de atributos, agora desagregados ao nível máximo, referente a cada habitação, e com um maior número de casos.

V.1 – Dados disponíveis (portal CASA SAPO)

Do total de 56571 registos de habitações publicadas no portal no período de 2001 a Fevereiro de 2010 para os municípios de Aveiro e Ílhavo, foram recolhidos 19969 referentes a habitações.

Os critérios de limpeza que levaram a esta redução basearam-se na escolha de imóveis:

- ✓ Publicitados, exclusivamente, para venda: como vimos anteriormente o tipo de negócio é um factor determinante das unidades de medida no indicador de preço, tornando obrigatório um processo de conversão e a utilização de variáveis de controlo que permitam distinguir os aspectos específicos do mercado para ambas as realidades (é facilmente aceitável que o arrendamento e a venda têm lógicas de mercado distintas).
- ✓ Sem incoerências internas ao nível das variáveis disponibilizadas: nomeadamente, coerência entre atributos físicos (por exemplo, não foram seleccionadas habitações que referem uma tipologia T6 com 2 quartos ou outras incoerências semelhantes).
- ✓ Existência de informação sobre a zona onde se localiza o imóvel e da delimitação territorial administrativa (freguesia) em que se localiza: esta informação é essencial para a conexão da base de dados com outros dados, referentes a atributos de localização. A escolha desta variável espacial deve-se à ausência de uma referência mais desagregada.
- ✓ Sem valores omissos: dado o número elevado de dados, considerou-se que a simples eliminação destes casos não afectaria substancialmente a construção de modelos representativos das realidades que se pretendem estudar. Por outro lado, a imputação destes valores, levariam a um aumento da complexidade do trabalho aqui apresentado, o que é inviável dentro dos constrangimentos de tempo disponíveis.

Como resultado, os dados recolhidos do portal Casa Sapo englobam os seguintes atributos, distribuídos por três categorias:

- **Atributos físicos básicos:** correspondem aos atributos que representam à informação básica para publicitação dos imóveis (ex: PREÇO, ÁREA, TIPOLOGIA, PRESERVAÇÃO)

- **Atributos físicos descritivos:** extraídos a partir de um curto texto comercial que acompanha o anúncio dos imóveis no portal com objectivos comerciais declarados (*marketing*). Referem-se a aspectos que o vendedor considera chave para prender a atenção do utilizador do portal. A extracção destes atributos foi realizada no âmbito do trabalho de Marques et al (2010) e envolveu a realização dos seguintes procedimentos:

  - Identificação do conjunto de palavras-chaves (ex: *varanda*, *lareira*, *duplex*, etc.) relevantes na caracterização física de uma habitação.

  - Determinação (automática) da existência ou não de cada uma destas palavras no texto associado a cada imóvel.

  - Armazenamento da informação em 13 variáveis binárias (variáveis tipo *dummy*, já descritas anteriormente), atribuindo o valor "1" se a referida palavra está contida no texto ou "0", caso contrário. A lista dos 13 atributos binários adicionados é apresentada na Tabela 7

Tabela 6 Variáveis tipo *dummy* (atributos binários) adicionados à base de dados

NOME DO ATRIBUTO BINÁRIO	DESCRIÇÃO
ATRIB_1_DUPLEX	alojamento do tipo duplex
ATRB_2_ARREC	existência de arrecadação
ATRB_3_VARANDA	existência de varanda
ATRB_4_SOTAO	existência de sótão
ATRB_5_TERRACO	existência de terraço
ATRB_6_LUGARGARAGEM	existência de lugar de garagem
ATRB_7_GARAGEM	existência de garagem
ATRB_8_PISO	piso em que se situa o alojamento
ATRB_9_AQUECIMENTO	existência de infraestrutura de aquecimento
ATRB_10_WC	existência de wc
ATRB_11_REMODELADO	existência de obras de remodelação
ATRB_12_LAREIRA	existência de lareira
ATRB_13_HIDROMASSAGEM	existência de hidromassagem

- **Atributos espaciais:** correspondem aos atributos que representam à informação de localização dos imóveis: *i*) CONCELHO, FREGUESIA - referenciando-se a delimitações administrativas; *ii*) ZONA[18] – uma referência classificatória ambígua, associada a delimitações históricas, a conjurações sociais e a critérios estabelecidos pelo agente da oferta ou seu intermediário; assim, não possui fronteiras geográficas efectivamente definidas.

[18] Este atributo será designado também como MICROZONA, de forma a distinguir de outros atributos que serão incorporados e que serão descritos mais à frente.

V.2 – Dados disponíveis (portal SAPO MAPAS)

Os dados disponibilizados pelo portal Sapo Mapas referem-se a pontos no espaço que determinam a localização de determinadas funções (equipamentos, serviços e locais de interesse para a população residente e pendular / visitante). O tratamento desta informação tem sido alvo de múltiplas abordagens, como comprovam os trabalhos de Li et al (1980), de Kiel et al (2008) ou de Sucahyono (2006).

A forma mais expedia de tratar esta informação baseia-se na construção de novas variáveis que traduzam características de vizinhança ou localização ancoradas na medida de distância. Criticamente, autores como Ross et al (2009) referem que as relações matemáticas dadas pela escolha de quaisquer dois pontos, numa grelha cartesiana, resulta numa grande sensibilidade das variáveis distância "a" (ou "de"). Neste contexto os autores argumentam que os coeficientes do modelo de regressão são altamente sensíveis, registando-se a inclusão ou exclusão inconsistente de pontos e uma incorrecta determinação dos respectivos coeficientes. Para contrabalançar estes problemas, sugerem a utilização de dados mais agregados, embora não apresentem alternativas metodológicas para esta abordagem.

Em termos espaciais, os dados disponíveis no portal Sapo Mapas estão desagregados ao maior nível possível: cada função encontra-se representada por um ponto no espaço. Já a informação disponível nos dados do portal Casa Sapo, apresenta uma única variável espacial: de natureza categórica e não georreferenciada – a MICROZONA. No entanto, empiricamente verifica-se que a determinação dessa localização pode ser obtida pela aproximação a denominações de delimitações existentes noutras fontes de informação: designação histórica de áreas territoriais não delimitadas; classificação empírica dos profissionais do sector imobiliário – na sequência da distinção de áreas territoriais comerciais, com dinâmicas diferenciadas; designações atribuídas por departamentos da administração pública, nomeadamente, em departamentos que incidem a sua actividade no ordenamento do território e concretamente no desenho urbano.

O conhecimento empírico de muitas destas designações, apoiado em sistemas de informação geográfica, permite estabelecer localizações aproximadas para a localização no espaço do centróide representativo de cada uma das zonas referidas na base de dados

Casa Sapo. Este é um processo manual, realizado com o auxílio de um *software* de manipulação de sistemas de informação geográfica – SIG[19], e o resultado final apresenta-se na Figura 9.

Com a colaboração da equipa LabSapo[20] (Laboratório Sapo da Universidade de Aveiro) foi possível obter as coordenadas de um conjunto de *pontos de interesse*[21] para os municípios de Aveiro e Ílhavo.

Os dados encontram-se catalogados em 11 categorias num total de 975 pontos:

Áreas verdes (41)
Desporto (81)
Mobilidade (30)
Utilidades (169)
Comércio (53)
Divertimentos (217)
Serviços de saúde (53)
Áreas industriais (7)
Cultura (21)
Educação (138)
Elementos turísticos (165)

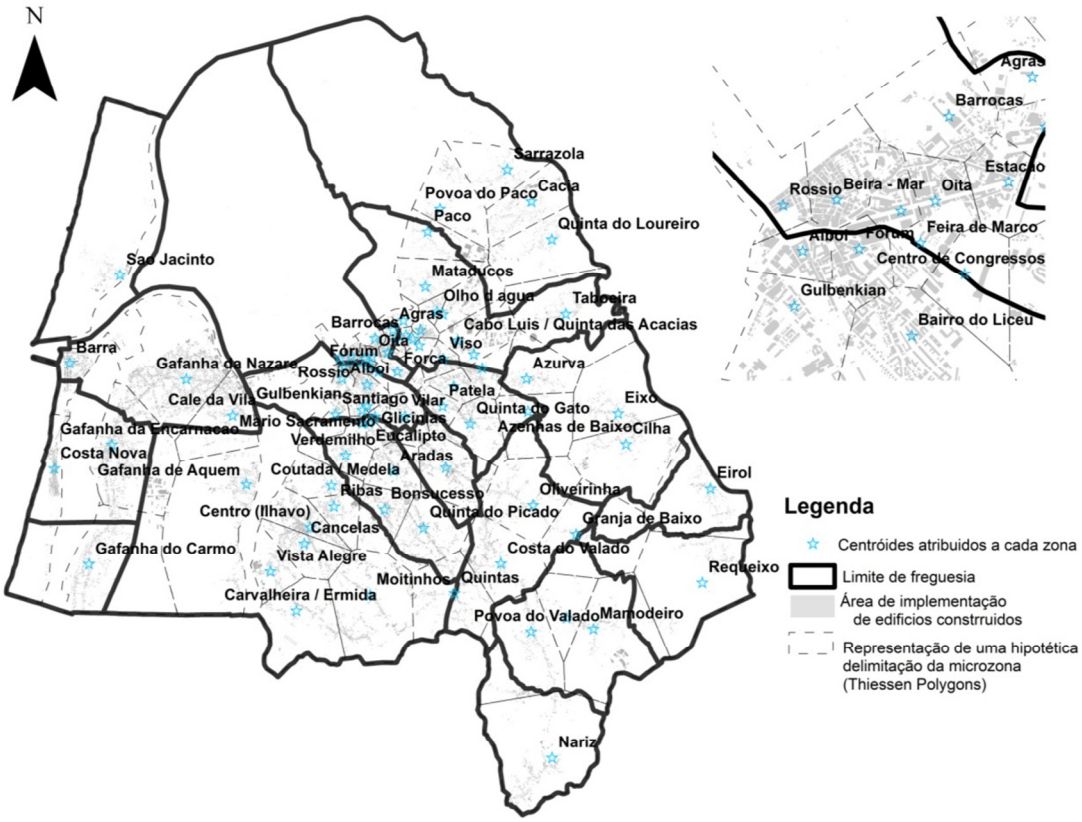

Figura 9 Representação espacial dos centróides georreferenciados de cada uma das zonas mencionadas no atributo zona da base de dados do portal Casa Sapo. Apresenta-se ainda a delimitação obtida pela aplicação da função geométrica para estimação dos limites de cada zona e os limites administrativos oficiais das freguesias (dos municípios de Aveiro e Ílhavo

[19] Software ArcGIS 9.3, da empresa ESRI. Para efeitos de representação gráfica, estimaram-se delimitações das zonas a partir da ferramenta *Thiessen Polygons*. Mais informações em http://www.esri.com/software/arcgis/index.html

[20] Mais informações em http://labs.sapo.pt/ua/

[21] São designados pela plataforma Sapo Mapas por *pontos de interesse* o conjunto de *layers* relativos à identificação de serviços de interesse geral: comércio, equipamentos, serviços (correios, instituições bancárias, …) e elementos culturais / paisagísticos / lúdicos (igrejas, conventos, …)

V.2.1. Definição de atributos de localização

É possível delimitar, empiricamente, áreas territoriais que traduzem maior densidade de pontos. Estas áreas são tradicionalmente associadas aos centros urbanos ou a conceitos como o de CBD (*Central Business Disctric*) que, neste caso, constitui uma boa opção uma vez que o conceito refere-se à área territorial de maior densidade de serviços.

A utilização de centralidades é comum na literatura para definição de atributos de localização associados à habitação. A incorporação destes atributos é frequentemente efectuada a partir das distâncias ou, em alternativa, pela introdução de variáveis classificatórias (do tipo *dummy* já definidas anteriormente).

No presente caso de estudo, definiram-se três centralidades territoriais. Duas são estabelecidas pela mancha de pontos de interesse e referem-se a uma hipotética delimitação do centro de Aveiro e do centro de Ílhavo. A terceira categoria refere-se às *praias*. A inclusão das praias nesta classificação deve-se à relevância do elemento *praia* no lazer generalizado da população. Embora o efeito centralidade seja fundamentalmente sazonal, várias tendências socioeconómicas apontam para a maior atractividade das áreas costeiras na concentração de população (residente e temporária) Esta centralidade é intrínseca ao território onde se localiza, não sendo possível deslocaliza-la, ao contrário de um centro funcional, que é estabelecido a partir de dinâmicas socioeconómicas.

Assim, as manchas que delimitam estas centralidades, desenhadas apenas a partir da avaliação visual, sobrepõe-se às seguintes MICROZONAS (ver Figura 10):

Centro de Aveiro: constituído pelas MICROZONAS Alboi, Avenida Dr Lourenco Peixinho, Bairro de Santiago, Bairro do Liceu, Barrocas, Beira Mar, Carramona, Centro de Congressos, Escolas, Esgueira, Estacão, Eucalipto, Feira de Marco, Forca, Fórum, Glicínias, Gulbenkian, Mário Sacramento, Oita, Quinta do Cruzeiro, Rossio, Santiago, Viaduto, Vila – Jovem / Santiago

Centro de Ílhavo: constituído pelas MICROZONAS Cancela, Centro (Ílhavo)

Praias: constituída pelas MICROZONAS Barra, Costa Nova, São Jacinto

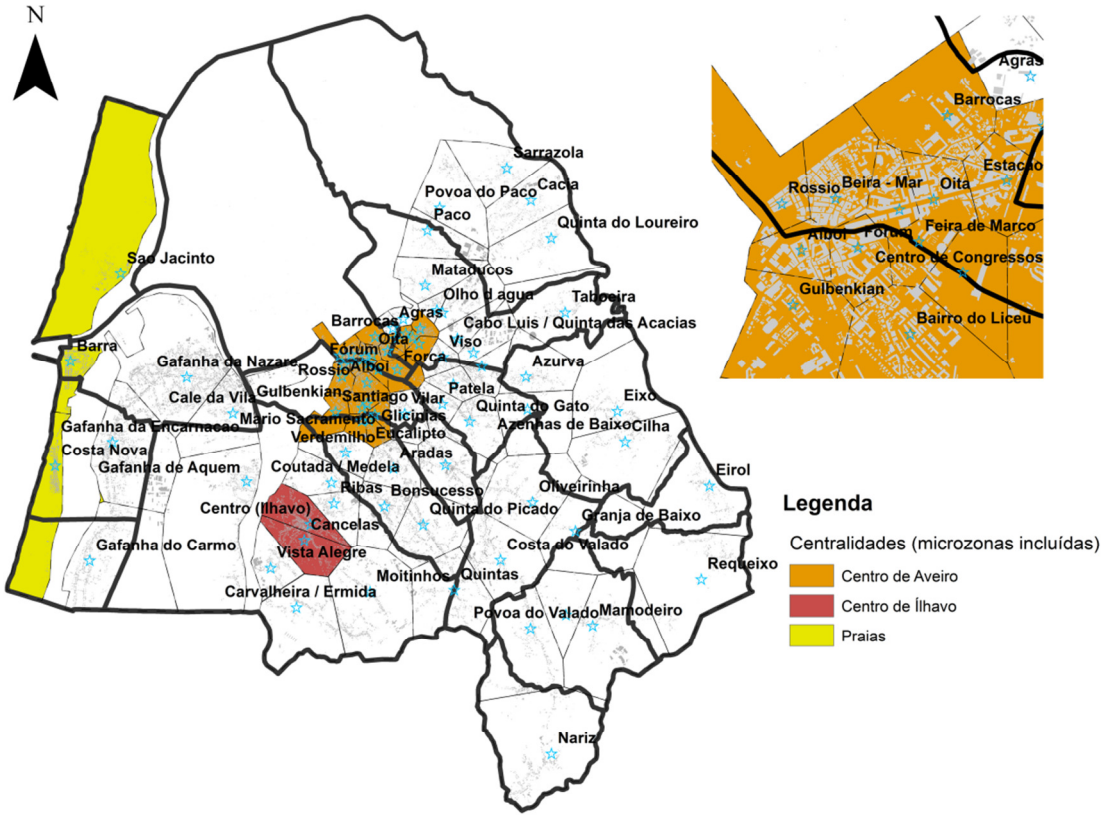

Figura 10 Representação das delimitações das centralidades definidas visualmente a partir dos limites hipotéticos das macrozonas incluídas.

V.2.2. Definição de atributos de vizinhança

A Figura 11 representa a distribuição espacial dos *pontos de interesse* dos concelhos de Aveiro e Ílhavo. A informação georreferenciada provém da base de dados do serviço Sapo Mapas[22].

Os dados disponibilizados permitem desenvolver indicadores que traduzam diferenças da vizinhança de cada uma das zonas. Cada zona tem características específicas que lhe podemos associar, determinadas pelo tipo e quantidade de *pontos de interesse* que se distribuem na sua envolvente.

[22] Informação recolhida e disponibilizada pelo LabSapo

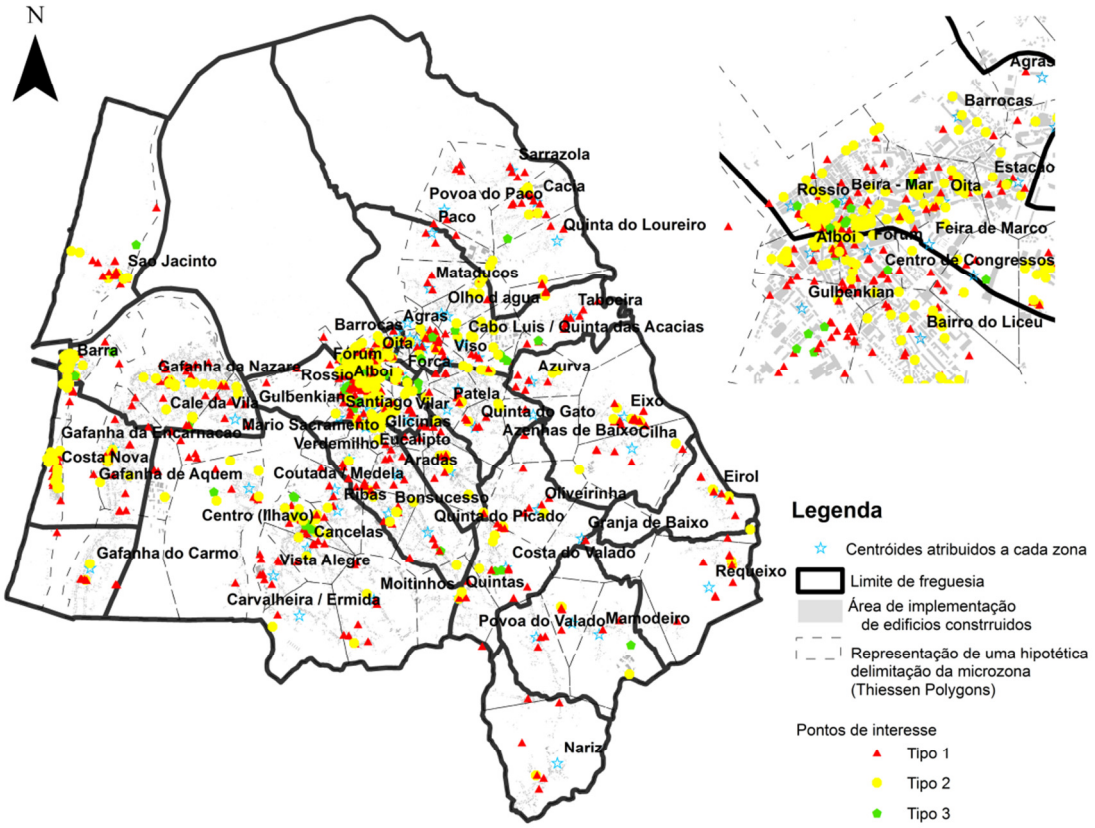

Figura 11 Pontos de interesse das várias categorias disponibilizadas no portal Sapo Mapas. Apresenta-se a sua distribuição espacial com a subdivisão categórica do tipo 1, 2 e 3.

É crível que os *pontos de interesse* exerçam um efeito de vizinhança, espacialmente limitado, a cada habitação. Podemos considerar como uma boa hipótese de trabalho que as funções (disponibilizadas pela localização de equipamentos e serviços) constituem atributos locais, se localizados dentro de uma dada delimitação territorial.

Não existindo um limite associado à variável zona, propõe-se como uma aproximação teórica (dado que a correcta delimitação implica uma investigação autónoma, de grande complexidade, que não cabe neste trabalho realizar) a utilização de conhecimento empírico, associado ao desenho urbano – por ser uma área técnico – científica que envolve o desenho do espaço ao nível local e, assim, com potencial para fornecer importantes critérios de delimitação da vizinhança. O princípio utilizado neste trabalho baseia-se no facto de a vizinhança poder ser determinada pelos raios estabelecidos para uma deslocação pedonal confortável. É comum a utilização, em desenho urbano, de raios de 600m para o alcance máximo confortável, quando

deslocação efectuada a pé e de forma muito frequente. Raios de alcance maiores correspondem a distâncias que apenas são percorridas pelo peão por motivos esporádicos.

Parte-se da hipótese empírica de que é aceitável definir a vizinhança de uma habitação como a organização percepcionada pelo residente (ou potencial residente) das funções proporcionadas pelo espaço envolvente. Considerando que a percepção do espaço é obtida através do contacto físico, é razoável admitir que a deslocação pedonal do residente fornece a percepção global do que o rodeia e que esse é um aspecto psicossocial importante na definição do conceito de vizinhança.

As deslocações, para aceder ao conjunto de funções que povoam a vizinhança de uma habitação, são efectuadas consoante o grau de relevância, da função, para o dia-a-dia do residente – isto, considerando exclusivamente deslocações pedonais. Por exemplo, sem entrar em considerações de ordem psicossociológica, pode-se admitir que a deslocação a uma padaria é essencial para a maioria da população, efectuando-se todos os dias. Caso a função esteja a uma distância confortável, é crível que essa deslocação seja efectuada a pé, o que proporciona ao residente a associação da padaria à vizinhança da sua habitação. Se tiver de efectuar uma deslocação num meio de transporte alternativo, o esforço maior que representa esse tipo de deslocação para o residente e o afastamento natural que provoca do espaço físico, pode significar que a padaria passa a ser um mero serviço, acessível, mas não parte de uma característica de vizinhança.

Na ausência de uma metodologia amplamente estabelecida para tratar este problema, opta-se por esta formulação, baseada no senso comum e fundamentada por alguns conhecimentos associados ao estudo do espaço e do território, adquiridos ao longo do percurso académico que efectuei. É claro que esta é uma suposição meramente empírica e facilmente questionável. Contudo, as limitações de tempo e recursos obrigam a não aprofundar esta questão.

Estabelecidos os princípios, apresenta-se o conceito de vizinhança como o alcance máximo, com conforto, para deslocações locais, pedonais, dos residentes numa dada MICROZONA, com objectivos de usufruir de funções existentes no espaço envolvente.

É necessário estabelecer uma grelha de classificação da relevância de cada função. Mais uma vez, parte-se de uma avaliação *ad-hoc,* determinada pela expectável

intensidade de uso e importância, para a população em geral dos *diferentes pontos de interesse*. Para construir esta hipótese de trabalho consideram-se as categorias compiladas do serviço Sapo Mapas como diferentes tipos de funções, distribuídas de forma não homogénea pelo território. Cada conjunto destas categorias é repartido, utilizando-se como critério a classificação empírica dessa função, por três níveis de relevância considerados (ver Figura 11):

- **[T1] – Tipo 1**: Engloba os serviços e equipamentos dimensionados para um servir a população num alcance limitado. São funções com uma intensidade de uso quase diário (ou pelo menos semanal). Apontam-se como exemplos o comércio de proximidade, associado à alimentação: *mercearias*, *talhos*, *mercados*, *padaria*, etc. A consideração destas funções para as características de vizinhança é determinada por um raio de 600 *m*, em linha recta, calculado a partir da localização espacial da habitação.

- **[T2] – Tipo 2**: Conjunto de funções com uma utilização esporádica (ou direccionadas para servir um número reduzido de residentes) definidas pela intensidade de utilização *baixa*/*casual*. Neste tipo de funções, é expectável que o residente esteja disposto a caminhar mais alguns metros para alcançar esta função, pelo que se estabeleceu o alcance, dado pelo raio de 1200 *m*.

- **[T3] – Tipo 3**: Conjuntos de funções que têm uma necessária localização espacial, muito embora sejam utilizados *muito excepcionalmente* pelos residentes numa determinada zona. São equipamentos e serviços que servem uma população supra-local e, muitas vezes, inclusivamente, populações externas aos municípios de Aveiro e Ílhavo. Podemos referir como exemplo, o *estádio de futebol*, o *hospital* e o *tribunal*. No entanto, é interiorizado pelos residentes como parte da vizinhança da sua habitação, embora, neste caso, permitam abarcar na sua vizinhança um raio mais lato, que se definiu neste trabalho de 1800 *m*.

As 11 categorias disponíveis são, com este critério, transformadas num grupo de 33 funções, através do desdobramento classificatório dos *pontos de interesse* de uma dada categoria por cada um dos três tipos de relevância atrás estabelecidos. Por exemplo, a categoria COMÉRCIO foi subdividida em três tipos de funções: COMÉRCIO TIPO 1, COMÉRCIO TIPO 2 e comércio tipo 3. O mesmo processo foi implementado para

as restantes categorias, podendo existir categorias que não englobem conjuntos em todos os tipos de relevância.

Assume-se também que a localização do centróide da zona é uma aproximação à localização real das habitações (uma aproximação grosseira mas aceitável, dadas as limitações associadas aos dados espaciais de que dispomos). Para cada centróide, apenas os serviços e equipamentos que se encontram nos raios de alcance definidos são efectivamente relevantes para caracterizar a vizinhança de uma habitação.

O facto de a distribuição das MICROZONAS não ser homogénea no território em estudo terá como consequência que os limites de alcance, determinados a partir do centróide da MICROZONAS, podem, em muitos casos e para centróides com localizações muito próximas, sobrepor-se. Este não é um problema relevante, uma vez que o conceito de vizinhança que definimos permite uma necessária partilha.

Para aprofundar a diferenciação da vizinhança associada a cada MICROZONA, introduziu-se uma nova medida, a *distância mínima*, calculada a partir do centróide da MICROZONA, em relação a um qualquer ponto no espaço, do conjunto de *pontos de interesse* de uma dada função, que se encontre no círculo de alcance definido pelo tipo de relevância a que se refere a função.

A partir da construção conceptual anterior, podemos definir o conceito de vizinhança de uma MICROZONA através de duas medidas:

i) a quantidade de *pontos de interesse* de uma dada função que se encontram dentro do círculo determinado pelo alcance determinado pelo tipo de relevância da função a que se refere;

ii) a distância mínima determinada a partir do centróide, ao *ponto de interesse* menos distante do conjunto de pontos de cada função.

Para determinar uma medida que sintetize estes dois tipos de características, seleccionou-se a aproximação descrita por Lopes (2010), inspirada no conceito de *potencial* da física clássica e adaptada para o problema em estudo. Assim, o potencial

produzido pelos *pontos de interesse* pode ser determinado por uma função[23] *Pot (C)* tal que:

$$Pot(C)_z = \frac{\sum_{i=1}^{a} C_i}{\min(d_{zc})} \qquad [16]$$

sendo,

- ✓ *Pot* (*C*) – o potencial no centróide da MICROZONA z para uma dada função *C*
- ✓ C_i – os *pontos de interesse* do conjunto de pontos da função *C*,
- ✓ a – número total de *pontos de interesse* que num circulo de raio definido pelo tipo de relevância associado aos pontos da função C
- ✓ d_{zc} - a distância do centróide da MICROZONA *z* a cada um dos pontos de interesse C_i

Figura 12 Representação espacial dos critérios para caracterização da vizinhança dada pelos equipamento s de educação de tipo 3 (símbolo verde). O cálculo do potencial baseia-se na utilização da distância mínima do centróide de cada MICROZONAS ao ponto de interesse (símbolo verde) e no número de pontos de interesse que ocorrem dentro dos limites definidos pelo hipotético circulo dado pelo alcance de influência de um dado ponto de interesse em relação ao centróide da microzona.

Na Figura 12 demonstra-se espacialmente como são determinados e medidos os critérios que permitem definir as características de vizinhança de uma dada MICROZONA. Como se depreende da equação [16], o potencial depende da distância ao ponto de

[23] Não esquecer que a menção de função refere-se, aqui, ao desdobramento das categorias de *pontos de interesse* por tipos de relevância dessas mesmas categorias - por exemplo, equipamentos ou serviços, representados por pontos no espaço, que prestam uma função de comércio tipo 1.

interesse mais próximo calculada a partir do centróide e ainda da massa, correspondente ao número de pontos de interesse abarcado no círculo hipotético do alcance considerado (no exemplo, para um serviço / equipamento tipo 3, representa um raio de alcance de 1800 m). Como se pode ver, apenas Gafanha da Nazaré e Cale da Vila têm o ponto de interesse dentro do limite circular; neste caso, o desempate é determinado pela distância. Como se pode constatar na Figura 13, apesar de diferenças de distância pequenas, a MICROZONA Gafanha da Nazaré, situando-se mais perto do referido ponto de interesse, caracteriza-se por uma maior medida de vizinhança (potencial).

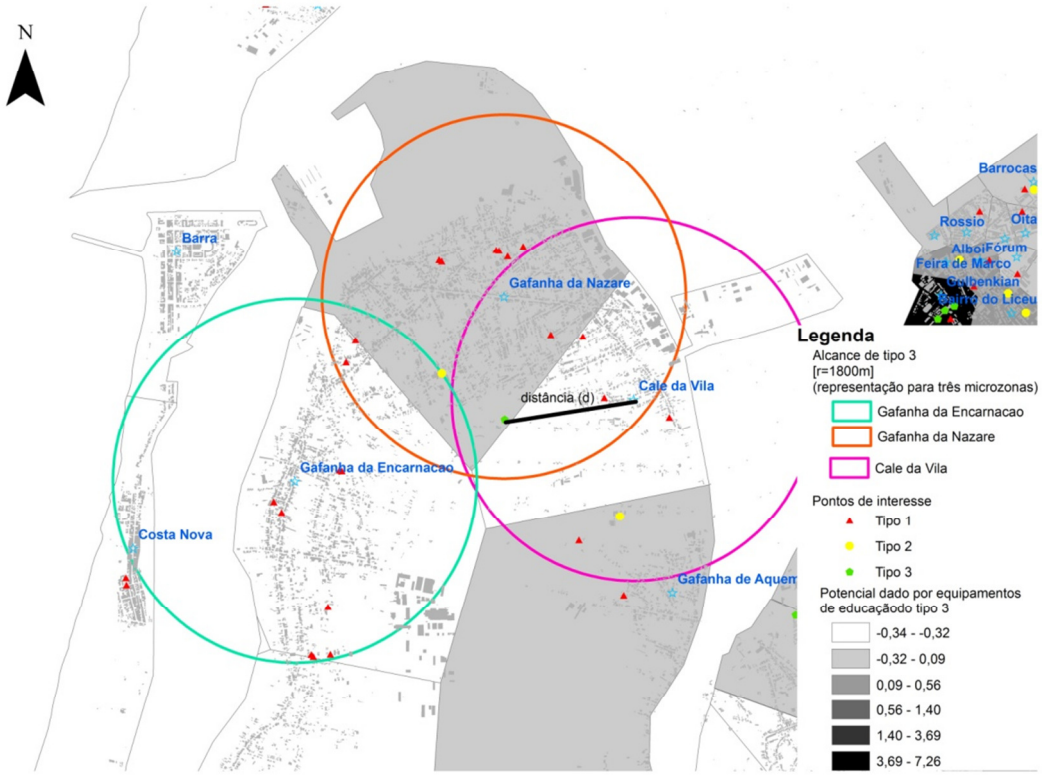

Figura 13 Resultados do potencial determinado pelos equipamentos educativos do tipo 3, enquanto elemento caracterizador da vizinhança das diferentes microzonas.

O potencial é um indicador sem qualquer pretensão dimensional, procurando simplesmente oferecer uma ferramenta de diferenciação das vizinhanças associadas a cada MICROZONA.

V.3 – Constrangimentos associados às variáveis recolhidas

V.3.1. Limitações gerais

A função comercial do portal Casa Sapo baseia-se no fornecimento de uma ferramenta complementar de mediação imobiliária. Complementando a função de vendedores e compradores de produtos imobiliários, o portal é alimentado:

- ✓ maioritariamente por empresas de mediação imobiliária, para as quais a empresa oferece também uma solução de software de gestão imobiliária – *ImoGuia* – que permite a publicitação automática dos imóveis no portal.
- ✓ em menor escala, directamente pelos proprietários. Neste caso, a interface de alimentação é assegurada por um mecanismo de exportação associado à solução de software na versão gratuita ou através de preenchimento de formulário online.

A informação de venda do imóvel é disponibilizada na plataforma Casa Sapo e visualizada de forma gratuita por qualquer potencial comprador.

A informação associada à cadeia comercial de mediação tem limitações inerentes:

- ✓ a ausência de exclusividade na comercialização de imóveis e a não identificação do proprietário – por motivos de salvaguarda da privacidade, segredo de negócio, entre outros, traduz-se na possibilidade de o imóvel surgir repetidas vezes publicado na plataforma.
- ✓ a publicação de um imóvel no portal Casa Sapo tem um objectivo comercial que pode levar o vendedor a omitir ou alterar atributos que potenciem a sua desvalorização.
- ✓ a publicação de informação relativa à maioria dos atributos no portal, a partir das diferentes interfaces, não é efectuada através de campos fechados ou relacionais, tornando possível a existência de valores incoerentes.
- ✓ a possibilidade de utilizar um campo descrição de texto livre permite a adição de informação com potencial interesse. A empresa não tem implementado

esquemas semi-automatizados ou automatizados para realizar o seu tratamento.

A disponibilidade de informação proveniente deste campo provém de um mecanismo de identificação de palavras-chave, no campo descrição de cada imóvel, realizado numa perspectiva exploratória.

V.3.2. Limitações do indicador de preço

O sentido unidireccional do fluxo de informação, que alimenta o portal, é responsável pelas especificidades da variável referente ao valor do imóvel. Ao contrário do proposto nos modelos econométricos – da utilização dos preços de transacção – a variável disponibilizada refere-se à avaliação monetária proposta pelo proprietário da habitação.

Assumindo os constrangimentos associados ao mercado imobiliário, que têm vindo a ser discutidos nos capítulos anteriores, podemos considerar que, no curto prazo, a oferta é limitada, tendencialmente constante. Desta forma, os agentes da oferta são os principais responsáveis pelo real preço de transacção – o que nos permite considerar o indicador de preço disponível como uma boa aproximação do preço de transacção real.

O portal é maioritariamente alimentado por empresas de mediação imobiliária. O seu conhecimento profundo do mercado permite aferir ainda maior consistência ao pressuposto de que os valores não variam significativamente dos reais valores de transacção.

V.3.3. Limitações dos indicadores territoriais

Os indicadores territoriais associados a cada imóvel são limitados. Existe uma única variável, referente a uma delimitação territorial que designamos por MICROZONA. Como principais limitações associadas a esta variável, destaca-se:

- ✓ inexistência de delimitação geográfica / espacial;

- ✓ campo proveniente de um mecanismo de resposta aberta, onde a correcta classificação de uma habitação depende da percepção empírica, efectuada por cada utilizador da plataforma;
- ✓ não existe uma relação directa entre a variável classificatória MICROZONA e as variáveis que referem limites administrativos FREGUESIA / MUNICÍPIO, dificultando ainda mais a percepção dos limites territoriais.

V.4. – Implementação do processo de data mining

O processo de *RapidMiner* implementado seguiu o desenho proposto na secção III.6.2, embora neste caso de estudo existam pormenores importantes, específicos do nível de detalhe e da quantidade e qualidade dos dados.

Descrevem-se as tarefas implementadas em cada fase do processo, de acordo com a metodologia adoptada neste trabalho:

Objectivo: identificar os atributos determinantes no processo de formação do preço da habitação

Pré-processamento: embora seja provável que o *RapidMiner* permita implementar todos os processos de limpeza e transformação descritos anteriormente, por uma questão de simplicidade e rapidez, optou-se pela sua implementação usando o programa *Microsoft Excel,* onde o domínio das ferramentas de manipulação de dados disponíveis é maior. Executaram-se as seguintes tarefas de transformação de variáveis:

- construção das variáveis de vizinhança, utilizando a metodologia proposta anteriormente
- integração das variáveis de vizinhança na base de dados global
- conversão das variáveis nominais para a sua representação numérica (requisito das técnicas de modelação utilizadas)
- recodificação da variável zona em 3 novas variáveis (do tipo *dummy*): MACROZONA CENTRO DE AVEIRO, MACROZONA CENTRO DE ÍLHAVO E MACROZONA PRAIAS.

Após a aplicação das transformações descritas, o projecto foi implementado em *RapidMiner* executando-se as restantes ferramentas, descrita na secção III.6.2.

Modelação: Nesta fase, similar aos estudos anteriores, o objectivo é construir um *modelo de preços hedónicos* baseado em modelos de regressão linear.

Avaliação: o volume grande de dados determinou a utilização de um esquema de validação *hold-out* uma vez que apresenta menor custo computacional.

Considerações específicas para a implementação

A reflexão teórica do primeiro capítulo fundamenta a aplicação de um *modelo de preços hedónicos* no contexto de um mercado, único. No entanto, a variável zona é definida na plataforma de publicitação de imóveis pela percepção dos utilizadores. Sendo um campo de resposta aberta e tendo a plataforma objectivos comerciais, é crível que a atribuição de uma zona possa associar algum facto intrínseco ao mercado que não está aqui a ser considerado. Neste aspecto, salientam-se as considerações de autores como Malpezzi (2008) e Palmquist (2005) que referem a possibilidade de um mercado poder subdividir-se em submercados. É o caso apresentado anteriormente para o tipo de negócio – *venda* vs *arrendamento* – que pode ser visto na perspectiva de dois submercados da habitação ou como um único, desde que identificados atributos que permitam reflectir ambas as realidades.

No caso em apreço, o atributo espacial ZONA, pode assim referir-se a submercados, delimitados pelos agentes imobiliários. A forma encontrada para limitar este problema, sem aumentar a complexidade do trabalho – que implicaria a investigação e delimitação de submercados espaciais – consistiu na utilização de partições estratificadas. Com este procedimento garante uma representatividade mínima das zonas e, dessa forma, a limitação das interacções no modelo global provocadas, por características específicas de submercados espaciais.

V.4.1. Descrição da base de dados

A base de dados resultante após da aplicação dos processos de transformação descritos anteriormente contém 19969 observações e 47 atributos. Destes 47 atributos, 27 são numéricos e 21 categóricos. A Tabela 7 apresenta o nome e uma breve descrição de cada atributo, assim como algumas estatísticas descritivas (*média*, *desvio padrão*, *valores mínimos* e *máximos*) associadas a cada um deles. Esta base de dados foi a utilizada como entrada nos diferentes projectos de *RapidMiner* implementados neste estudo e descritos na secção III.6.2.

Tabela 7 Descrição dos atributos incorporados na base de dados, utilizados para a construção dos diferentes *modelos de preços hedónicos*

CODIGO	DESCRIÇÃO	PAPEL NO MODELO	TIPO VAR	ESCALA MEDIDA	ESTATÍSTICAS			TIPO ATR
					MÉDIA E DESVIO PADRÃO	MÍN	MÁX	
ID		id	nominal	-	-	-	-	-
MICROZONA	MICROZONA RECODE COD	grupo	nominal	-	-	-	-	-
V00_PRECO_M2	Indicador do preço de venda da habitação	variável dependente	real	€ / m^2	δ = 1134,6 +/- 384,9	152,5	5.714,3	
V01_AREA	Área da habitação	variável independente	real	m^2	δ = 154,1 +/- 83,8	20,0	600,0	E
V02_MACROZONA CENTRO AVR COD	Habitação localizada no centro de Aveiro	variável independente	nominal	-	δ = 0,3 +/- 0,4	0	1	L
V03_MACROZONA CENTRO ILH COD	Habitação localizada no centro de Ílhavo	variável independente	nominal	-	δ = 0,1 +/- 0,3	0	1	L
V04_MACROZONA PRAIAS COD	Habitação localizada junto às praias	variável independente	nominal	-	δ = 0,1 +/- 0,3	0	1	L
V05_PRESERVACAO_NOVO	Preservação das habitações novas	variável independente	ordinal	-	δ = 0,8 +/- 0,8	0	2	E
V06_PRESERVACAO_USADO_CONSERVACAO&IDADE	Preservação das habitações usadas	variável independente	ordinal	-	δ = 1,0 +/- 1,2	0	5	E
V07_TIPO_APART	Tipo de apartamento	variável independente	ordinal	-	δ = 1,4 +/- 0,9	0	2	E
V08_TIPO_MORADIA	Tipo de moradia	variável independente	ordinal	-	δ = 0,6 +/- 1,1	0	6	E
V09_ATRIB_1_DUPLEX	Referência a tipo de apartamento *duplex* no campo descrição	variável independente	nominal	-	δ = 0,1 +/- 0,3	0	1	E

V10_ATRB_2_ARREC	Referência a existência de espaço de arrecadação no campo descrição	variável independente	nominal	-	δ = 0,0 +/- 0,1	0	1	E
V11_ATRB_3_VARANDA	Referência a existência de varanda no campo descrição	variável independente	nominal	-	δ = 0,3 +/- 0,5	0	1	E
V12_ATRB_4_SOTAO	Referência a existência de sótão no campo descrição	variável independente	nominal	-	δ = 0,0 +/- 0,2	0	1	E
V13_ATRB_5_TERRACO	Referência a existência de terraço no campo descrição	variável independente	nominal	-	δ = 0,2 +/- 0,4	0	1	E
V14_ATRB_6_LUGARGARAGEM	Referência a existência de lugar de garagem no campo descrição	variável independente	nominal	-	δ = 0,1 +/- 0,4	0	1	E
V15_ATRB_7_GARAGEM	Referência a existência de garagem no campo descrição	variável independente	nominal	-	δ = 0,6 +/- 0,5	0	1	E
V16_ATRB_8_PISO	Referência ao piso no campo descrição	variável independente	nominal	-	δ = 0,0 +/- 0,2	0	1	E
V17_ATRB_9_AQUECIMENTO	Referência a existência de sistema de aquecimento central no campo descrição	variável independente	nominal	-	δ = 0,4 +/- 0,5	0	1	E
V18_ATRB_10_WC	Referência a existência de wc (ou a características específicas do wc) no campo descrição	variável independente	nominal	-	δ = 0,3 +/- 0,4	0	1	E
V19_ATRB_11_REMODELADO	Referência a tipo de preservação remodelado no campo descrição	variável independente	nominal	-	δ = 0,0 +/- 0,1	0	1	E
V20_ATRB_12_LAREIRA	Referência a existência de lareira no campo descrição	variável independente	nominal	-	δ = 0,2 +/- 0,4	0	1	E
V21_ATRB_13_HIDROMASSAGEM	Referência a existência de hidromassagem no campo descrição	variável independente	nominal	-	δ = 0,1 +/- 0,3	0	1	E
V22_POT_AMENIDADE_COMERCIO_T1	Área territorial com vizinhança funcional relativa a actividades de comércio com relevância tipo 1	variável independente	real	-	δ = 67,6 +/- 195,7	0,0	1350,6	V
V23_POT_AMENIDADE_COMERCIO_T2	Área territorial com vizinhança funcional relativa a actividades de comércio com relevância tipo 2	variável independente	real	-	δ = 44,5 +/- 180,3	0,0	3337,1	V
V24_POT_AMENIDADE_COMERCIO_T3	Área territorial com vizinhança funcional relativa a actividades de comércio com relevância tipo 3	variável independente	real	-	δ = 7,5 +/- 25,3	0,0	169,7	V
V25_POT_AMENIDADE_CULTURA_T1	Área territorial com vizinhança funcional relativa a equipamentos de cultura com relevância tipo 1	variável independente	real	-	δ = 33,4 +/- 111,0	0,0	1296,8	V
V26_POT_AMENIDADE_CULTURA_T2	Área territorial com vizinhança funcional relativa a equipamentos de cultura com relevância tipo 2	variável independente	real	-	δ = 21,2 +/- 76,8	0,0	685,2	V
V27_POT_AMENIDADE_CULTURA_T3	Área territorial com vizinhança funcional relativa a equipamentos de cultura com relevância tipo 3	variável independente	real	-	δ = 1,8 +/- 4,9	0,0	50,5	V

V28_POT_AMENIDADE_DESPORTO_T1	Área territorial com vizinhança funcional relativa a equipamentos de desporto com relevância tipo 1	variável independente	real	-	δ = 37,0 +/- 147,3	0,0	1363,9	V
V29_POT_AMENIDADE_DESPORTO_T2	Área territorial com vizinhança funcional relativa a equipamentos de desporto com relevância tipo 2	variável independente	real	-	δ = 36,6 +/- 166,2	0,0	1701,2	V
V30_POT_AMENIDADE_DESPORTO_T3	Área territorial com vizinhança funcional relativa a equipamentos de desporto com relevância tipo 3	variável independente	real	-	δ = 0,1 +/- 0,2	0,0	0,8	V
V31_POT_AMENIDADE_DIVERTIMENTOS_T2	Área territorial com vizinhança funcional relativa a actividades de divertimento com relevância tipo 2	variável independente	real	-	δ = 14826,6 +/- 44365,8	0,0	173472,3	V
V32_POT_AMENIDADE_DIVERTIMENTOS_T3	Área territorial com vizinhança funcional relativa a actividades de divertimento com relevância tipo 3	variável independente	real	-	δ = 128,5 +/- 621,2	0,0	5759,8	V
V33_POT_AMENIDADE_EDUCACAO_T1	Área territorial com vizinhança funcional relativa a equipamentos de educação com relevância tipo 1	variável independente	real	-	δ = 129,4 +/- 268,1	0,0	2461,0	V
V34_POT_AMENIDADE_EDUCACAO_T2	Área territorial com vizinhança funcional relativa a equipamentos de educação com relevância tipo 2	variável independente	real	-	δ = 64,0 +/- 202,7	0,0	787,5	V
V35_POT_AMENIDADE_EDUCACAO_T3	Área territorial com vizinhança funcional relativa a equipamentos de educação com relevância tipo 3	variável independente	real	-	δ = 6,5 +/- 16,2	0,0	257,3	V
V36_POT_AMENIDADE_ELEMTURISTICOS_T3	Área territorial com vizinhança funcional relativa a elementos turísticos com relevância tipo 3	variável independente	real	-	δ = 1336,2 +/- 7769,7	0,0	71854,1	V
V37_POT_AMENIDADE_MOBILIDADE_T1	Área territorial com vizinhança funcional relativa a infraestruturas de mobilidade com relevância tipo 1	variável independente	real	-	δ = 4,5 +/- 24,2	0,0	842,1	V
V38_POT_AMENIDADE_MOBILIDADE_T2	Área territorial com vizinhança funcional relativa a infraestruturas de mobilidade com relevância tipo 2	variável independente	real	-	δ = 1,6 +/- 8,9	0,0	148,5	V
V39_POT_AMENIDADE_AREASVERDES_T1	Área territorial com vizinhança funcional relativa a áreas verdes com relevância tipo 1	variável independente	real	-	δ = 49,8 +/- 225,7	0,0	4239,8	V
V40_POT_AMENIDADE_SERVSAUDE_T1	Área territorial com vizinhança funcional relativa a equipamentos de saúde com relevância tipo 1	variável independente	real	-	δ = 41,2 +/- 82,9	0,0	2392,1	V
V41_POT_AMENIDADE_SERVSAUDE_T2	Área territorial com vizinhança funcional relativa a equipamentos de saúde com relevância tipo 2	variável independente	real	-	δ = 45,7 +/- 187,5	0,0	2232,5	V
V42_POT_AMENIDADE_SERVSAUDE_T3	Área territorial com vizinhança funcional relativa a equipamentos de saúde com relevância tipo 3	variável independente	real	-	δ = 2,6 +/- 9,0	0,0	41,5	V
V43_POT_AMENIDADE__ZI_T2	Área territorial com vizinhança funcional relativa a zonas industriais com relevância tipo 2	variável independente	real	-	δ = 0,3 +/- 0,9	0,0	5,1	V

V44_POT_AMENIDADE__ZI_T3	Área territorial com vizinhança funcional relativa a zonas industriais com relevância tipo 3	variável independente	real	-	δ = 0,3 +/- 0,6	0,0	2,7	V
V45_POT_AMENIDADE_UTILIDADES_T1	Área territorial com vizinhança funcional relativa a utilidades várias com relevância tipo 1	variável independente	real	-	δ = 89,1 +/- 495,4	0,0	5037,7	V
V46_POT_AMENIDADE_UTILIDADES_T2	Área territorial com vizinhança funcional relativa a utilidades várias com relevância tipo 2	variável independente	real	-	δ = 3778,3 +/- 8089,3	0,0	76658,5	V
V47_POT_AMENIDADE_UTILIDADES_T3	Área territorial com vizinhança funcional relativa a utilidades várias com relevância tipo 3	variável independente	real	-	δ = 0,811 +/- 2,603	0,0	34,3	V

V.5 – Resultados

V.5.1. Análise global

A Tabela 8 apresenta os resultados globais dos diferentes modelos de regressão linear obtidos a partir do processamento da base de dados descrita na secção anterior composta por 47 atributos e 19969 registos. O processo de limpeza implementado em *RapidMiner* envolveu a aplicação dos operadores <REMOVE DUPLICATES> e <REMOVE CORRELATED ATRIBUTES>. O primeiro operador reduz o número de casos da base de dados para 12997. O segundo operador removeu o atributo V34_POT_AMENIDADES_EDUCACAO_TIPO2, que se apresenta altamente correlacionado com pelo menos uma das restantes variáveis independentes. Os atributos resultantes após as operações de limpeza são 46.

O problema da selecção de atributos que efectivamente têm uma relação linear com o preço das habitações torna-se muito relevante neste conjunto de dados uma vez que temos um considerável número de atributos e um grande volume de dados.

O modelo base, construído com 46 atributos, apresenta uma capacidade explicativa de 54,5% do preço dos imóveis. É expectável que a capacidade explicativa nestes modelos micro seja de uma ordem de grandeza menor que na anterior uma vez que o fenómeno em estudo envolve necessariamente um maior número de atributos, aumentando o risco de não serem seleccionados atributos chave na explicação do preço. Por outro lado, a este nível de especificidade, as formulações dos modelos tornam-se mais sensíveis uma vez que o pressuposto da relação linear de cada atributo com o preço pode não ser válida para todo o conjunto de atributos – especialmente quando são utilizadas variáveis como as referentes aqui aos atributos de vizinhança, sem uma base teórica sólida e, dessa forma, necessários testes à verificação dos pressupostos que se devem cumprir.

Tabela 8 Resultados globais, dos diferentes *modelos de preços hedónicos,* para o caso de estudo à micro escala

	SEM SELECÇÃO	COM SELECÇÃO							
		Abordagem embutida		Redução da dimensionalidade		Abordagem Filtro		Técnicas de pesagem	
		Árvore de regressão [M5 Prime]	Algoritmo Greedy [estratégia forward e critério AIC]	Com novas variáveis	Com variáveis representativas dos conceitos	Heurística FSS [estratégia backward e critério CFS]	Heurística FSS [estratégia forward e critério CFS]	PCA *Weighting*	SVM *Weighting*
Modelos	**M0**	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
Variáveis seleccionadas	**46**	39	36	17	16	**8**	2	11	**3**
Variáveis significantes	**40**	39	35	15	11	**8**	2	10	**3**
Capacidade explicativa	**0,545**	0,545	0,544	0,437	0,422	**0,499**	0,372	0,283	**0,433**
Variação no número de variáveis	-	-15,2%	-21,7%	-63,0%	-65,2%	**-82,6%**	-95,7%	-76,1%	**-93,5%**
Variação da capacidade explicativa	-	0,0%	-0,2%	-19,8%	-22,6%	**-8,4%**	-31,7%	-48,1%	**-20,6%**

Em relação aos oito modelos resultantes da aplicação dos diferentes processos de selecção de atributos, podemos observar que a capacidade explicativa dos modelos varia entre os referidos 54,5% e os meros 28% obtidos com o esquema de pesagem que utiliza como pesos a primeira componente da ACP (PCAWEIGHTING).

Na generalidade, verifica-se que a utilização das técnicas de selecção de atributos permite diminuir o número de atributos que fazem parte do modelo de preços hedónicos, embora neste caso de estudo – e ao contrário do observado no nível *macro*, existam perdas relevantes na capacidade explicativa dos modelos.

V.5.2. Análise da capacidade explicativa e número de variáveis

Similar à análise realizada no estudo anterior (modelo *macro*), a avaliação dos resultados baseia-se no balanço (*trade-off*) entre a *capacidade explicativa* e o *número de variáveis* incluídas no modelo. Nestes modelos não existe uma relação directa entre ambos os critérios, o que traduz uma dificuldade acrescida na escolha inequívoca do modelo mais eficaz.

No caso deste estudo (modelo *micro*) destacam-se duas abordagens na selecção dos atributos: *i*) a selecção de tipo filtro com estratégia de busca *backward* e medida de qualidade CFS (modelo **M5**); *ii*) o esquema de pesagem com máquinas de suporte vectorial (SVM) (modelo **M8**). O modelo de preços hedónicos construído após do processo de selecção por filtro representa uma diminuição de apenas 8,4% da capacidade explicativa em relação ao modelo de base (de 54,5% foi reduzida para 49,9 %) e uma redução de 83% no número de variáveis necessárias para explicar o preço da habitação (das 46 variáveis iniciais foram seleccionadas apenas 8). Por outro lado, o modelo resultante após a aplicação de um esquema de pesagem com SVM oferece uma capacidade explicativa de 43% (-21% da capacidade explicativa do modelo de base) mas com a vantagem de este incluir apenas 3 variáveis, obtendo assim uma redução de 94% das 46 variáveis iniciais.

Dos restantes modelos, os resultados não são muito significativos uma vez que ou apresentam reduções muito grandes na capacidade explicativa ou a capacidade de selecção associada aos algoritmos implementados é mais reduzida.

Por fim, cabe salientar que embora apresentando um bom comportamento no estudo anterior (modelo *macro*), a redução de dimensionalidade não é aqui de todo vantajosa uma vez que leva à selecção de um número considerável de variáveis – 17, sendo a relação linear com o preço em 2 das 17, não significante, neste modelo. O decréscimo na capacidade explicativa também é relevante, situando-se na ordem dos 20%. Acresce que as novas variáveis representam 17 conceitos, para os quais é necessário efectuar uma interpretação. A análise dos *loadings* não permite uma associação clara de cada um dos novos conceitos a conceitos empíricos do investigador. Esta dificuldade deve-se a uma difícil identificação de qual dos conceitos cada uma das variáveis iniciais efectivamente contribui. Este facto pode dever-se a uma maior dispersão efectiva da informação latente nos dados iniciais pelas várias variáveis. Um processo interessante para refinar esta análise seria a possibilidade de aplicar técnicas de rotação às novas componentes (que não são mais que novos eixos ortogonais), o que permitiria extremar os *loadings* e, consequentemente, potenciar a identificação dos conceitos – no entanto, o *software* não disponibiliza uma forma expedita de realizar esta operação, sendo necessário desenhá-la. Devido às naturais limitações do analista e à

natureza dos objectivos deste trabalho não se efectuou este passo, que impunha o desenho de um algoritmo ou a utilização de outro tipo de *software*, em alternativa[24].

V.5.3. Análise das variáveis seleccionadas

A Tabela 9 apresenta, para cada uma das variáveis seleccionadas e em cada um dos modelos construídos, os respectivos *coeficientes estandardizados.*

A análise destes coeficientes permite-nos efectuar a análise comparativa *ad-hoc* já realizada para a análise *macro* de estudo anterior. Como se pode verificar na tabela, também aqui não existem diferenças na ordem determinada pelos coeficientes, em cada modelo. Sistematicamente, é a mesma variável independente – quando seleccionada previamente – que induz a maior variação da variável dependente.

Tabela 9 Quadro síntese dos coeficientes estandardizados de cada uma das variáveis incluídas nos *modelos de preços hedónicos* construídos

	SEM SELECÇÃO **MODELO BASE**	COM SELECÇÃO							
		Abordagem embutida		Redução da dimensionalidade		Abordagem Filtro		Técnicas de pesagem	
		Árvore de regressão [M5 Prime]	Algoritmo Greedy [estratégia *forward* e critério AIC]	Com novas variáveis	Com variáveis representativas dos conceitos	Heurística FSS [estratégia *backward* e critério CFS]	Heurística FSS [estratégia *forward* e critério CFS]	PCA *Weighting*	SVM *Weighting*
Modelos	**M0**	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
Capacidade explicativa	**0,545**	0,545	0,544	0,437	0,422	**0,499**	0,372	0,283	**0,433**
V01_AREA	**-0,504**	-0,503	-0,503			**-0,420**	-0,445	-0,489	**-0,480**
V02_MACROZONA CENTRO AVR COD	**0,149**	0,149	0,150		0,244			0,105	
V03_MACROZONA CENTRO ILH COD	**-0,019**	-0,021	-0,012						
V04_MACROZONA PRAIAS COD	**0,437**	0,437	0,433		0,428	**0,347**	0,380		**0,377**
V05_PRESERVACAO_NOVO	**0,061**	0,061	0,063						

[24] Não obstante, salienta-se que o *RapidMiner* possui mecanismos intuitivos de exportação que permitem utilizar os dados imediatamente com outro tipo de ferramentas. Por uma questão de gestão de tempo e dado que nada indicia que o modelo de redução de dimensionalidade possa efectivamente constituir uma das melhores alternativas, optou-se por não prosseguir com a análise neste campo.

V06_PRESERVACAO_USADO	**-0,191**	-0,191	-0,191		-0,269	**-0,281**			**-0,292**
V07_TIPOLOGIA_APART	**-0,097**	-0,098	-0,098		0,247			-0,036	
V08_TIPO_MORADIA	**0,064**	0,064	0,064					0,065	
V09_ATRIB_1_DUPLEX	**-0,017**	-0,017	-0,017		-0,095				
V10_ATRB_2_ARREC	**-0,005**				-0,015				
V11_ATRB_3_VARANDA	**-0,021**	-0,021	-0,022						
V12_ATRB_4_SOTAO	**0,000**								
V13_ATRB_5_TERRACO	**0,030**	0,030	0,030						
V14_ATRB_6_LUGARGARAGEM	**-0,010**								
V15_ATRB_7_GARAGEM	**0,048**	0,044	0,046		0,046				
V16_ATRB_8_PISO	**0,018**	0,018	0,018		0,012				
V17_ATRB_9_AQUECIMENTO	**0,043**	0,043	0,044						
V18_ATRB_10_WC	**-0,031**	-0,031	-0,031						
V19_ATRB_11_REMODELADO	**-0,035**	-0,035	-0,035						
V20_ATRB_12_LAREIRA	**-0,043**	-0,042	-0,042			**-0,043**			
V21_ATRB_13_HIDROMASSAGEM	**0,027**	0,027	0,027						
V22_POT_AMENIDADE_COMERCIO_T1	**0,044**	0,040	0,052		-0,002	**0,073**			
V23_POT_AMENIDADE_COMERCIO_T2	**-0,020**	-0,029						-0,197	
V24_POT_AMENIDADE_COMERCIO_T3	**0,018**	0,019	0,018						
V25_POT_AMENIDADE_CULTURA_T1	**-0,020**	-0,023	-0,029					-0,003	
V26_POT_AMENIDADE_CULTURA_T2	**-0,002**				-0,020				
V27_POT_AMENIDADE_CULTURA_T3	**0,036**	0,036	0,037						
V28_POT_AMENIDADE_DESPORTO_T1	**-0,024**	-0,023	-0,023						

V29_POT_AMENIDADE_DESPORTO_T2	**0,018**	0,022	0,028						
V30_POT_AMENIDADE_DESPORTO_T3	**-0,033**	-0,034	-0,034		-0,043				
V31_POT_AMENIDADE_DIVERTIMENTOS_T2	**0,057**	0,057	0,086		0,027	**0,144**			
V32_POT_AMENIDADE_DIVERTIMENTOS_T3	**-0,221**	-0,210	-0,236		-0,027	**0,110**		0,255	
V33_POT_AMENIDADE_EDUCACAO_T1	**-0,005**								
V35_POT_AMENIDADE_EDUCACAO_T3	**-0,023**	-0,026	-0,025						
V36_POT_AMENIDADE_ELEMTURISTICOS_T1	**-0,014**								
V37_POT_AMENIDADE_MOBILIDADE_T1	**-0,036**	-0,034	-0,041						
V38_POT_AMENIDADE_MOBILIDADE_T2	**0,094**	0,088	0,102		-0,001				
V39_POT_AMENIDADE_AREASVERDES_T1	**0,297**	0,294	0,299		0,148				
V40_POT_AMENIDADE_SERVSAUDE_T1	**-0,009**								
V41_POT_AMENIDADE_SERVSAUDE_T2	**-0,133**	-0,118	-0,059					-0,051	
V42_POT_AMENIDADE_SERVSAUDE_T3	**0,043**	0,038				**0,158**		0,057	
V43_POT_AMENIDADE_ZI_T2	**0,046**	0,047	0,047						
V44_POT_AMENIDADE_ZI_T3	**0,086**	0,084	0,088					0,070	
V45_POT_AMENIDADE_UTILIDADES_T1	**0,135**	0,117	0,136						
V46_POT_AMENIDADE_UTILIDADES_T2	**0,036**	0,030	0,030		0,012				
V47_POT_AMENIDADE_UTILIDADES_T3	**0,082**	0,077						0,140	

Assinalado com uma gradação de cor encontram-se as variáveis sistematicamente incluídas e significantes, o que se tomou como um indicador complementar da relevância efectiva de um dado atributo. Podemos identificar três níveis de importância:

- **NÍVEL 1 – Extremamente Relevante** 

  V01 e V04 [*área* e a localização *praias*]

- **NÍVEL 2 – Muito Relevantes** 

  V06 e V32: [*habitações usadas e respectivo grau de preservação* e *potencial de funções ligadas aos divertimentos do tipo 3* (funções de menor relevância / vizinhança alargada)]

- **NÍVEL 3 – Relevantes** 

  V02, V05 [localização no centro de Aveiro e tipologia de apartamento]

Os atributos que representam a *área da habitação* (V01) e a localização *praias* (V04), respectivamente, foram considerados extremamente relevantes. São sistematicamente seleccionados e incluídos nos diferentes modelos, surgindo em 7 do total de 9 modelos (coma ressalva de que o modelo que utiliza as componentes de uma ACP como novas variáveis não é directamente comparável com os restantes).

No conjunto de todos os modelos, em que surgem ambos os dois atributos, estes são os mais determinantes para a formação do preço das habitações. A *área* a surge em primeiro lugar, sendo que a variação positiva de um desvio padrão no atributo área representa uma variação negativa de 47% do desvio padrão, no preço. Já o atributo *praias* indica-nos que uma habitação, com as mesmas características, se estiver localizada numa das zonas referentes às praias, é expectável que o preço por metro quadrado seja 41% mais elevado que do desvio padrão do preço.

No segundo nível de relevância (*muitos relevantes*) surgem os dois atributos, V06 e V32, em 6 dos 8 modelos comparáveis. A variável V06 é estável em todos os modelos, uma vez que apresenta a mesma posição na ordem determinada pelos coeficientes *b*. O valor dos seus coeficientes estandardizados é, em média, - 0,236, não apresentando uma variação relevante de modelo para modelo. Tem, naturalmente, uma relação linear negativa com o preço uma vez que as categorias de preservação mais elevadas reflectem habitações em pior estado de conservação e de maior idade. A variável V32 (o único atributo de vizinhança) apresenta também instabilidade nos valores que apresenta. Os

coeficientes estandardizados apresentam uma grande variação nos valores absolutos (entre 0,255 e 0,110) a que acresce o facto de a sua relação linear ser negativa em 2/3 dos modelos e positiva no restante 1/3, o que não permite considerar com segurança este atributo a partir desta avaliação *ad-hoc*.

No último nível de relevância surge novamente um atributo estrutural, a tipologia de apartamento (V05), e o atributo centro de Aveiro (V02) que descreve as habitações localizadas no conjunto de zonas que caracterizam esta área. Destas variáveis, a V02 apresenta um coeficiente que conserva a ordem nos vários modelos em que é seleccionado, bem como não apresenta grande variação no coeficiente estandardizado, com um valor médio de 0,159

V.5.4. Análise dos níveis de significância

A Tabela 10 apresenta-nos os níveis de significância (para níveis de confiança de 95%) das variáveis incluídas em cada um dos 8 modelos (o modelo **M3**, relativo às componente, pelo facto de não ser directamente comparável e pelo facto de nos indicadores de relevância anteriores não apresentar reais mais valias, não é considerado nesta análise).

Verifica-se que neste caso de estudo não existe um número elevado de variáveis não significantes, para todos os modelos construídos. Este facto pode estar relacionado com o nível de detalhe com que estamos a trabalhar. Tal como descrito na reflexão teórica, a complexidade associada ao bem habitação (em termos individuais) é traduzida pelo grande número de atributos que podemos utilizar para a caracterizar. A identificação deste conjunto de atributos pelo investigador, à *priori*, não é uma tarefa fácil, sendo limitada pela informação disponível. Os níveis de significância apontam efectivamente que a maioria das variáveis consideradas possui uma relação linear com o preço, embora, uma grande parte dos atributos, apresente impactos diminutos – atributos onde a variação de uma unidade, reflecte-se numa variação inferior a 10% do desvio padrão do preço.

Os modelos que incluem maior quantidade relativa de atributos não significantes são obtidos pela técnica de redução baseada na redução de dimensionalidade (com

realce para um modelo **M4**). Este aspecto é um indicador de que, embora muitos dos atributos recolhidos possam ser considerados individualmente como características determinantes no preço da habitação, dos conceitos gerais que representam, obtidos pela ACP, é expectável que os atributos recolhidos englobem vários conceitos sem efectiva relação com o preço da habitação. Esta é um aspecto complementar para as baixas capacidades explicativas que encontramos nestes modelos micro: para além das questões de formulação, também uma quantidade significativa de atributos recolhidos pode não ser efectivamente muito relevante, pelo que não existe razão para aceitar modelos muito detalhados, no contexto do objectivo deste trabalho.

Tabela 10 Quadro síntese dos *p-values* associados a cada uma das variáveis para cada um dos modelos construídos

	SEM SELECÇÃO **MODELO BASE**	COM SELECÇÃO							
		Abordagem embutida		Redução da dimensionalidade		Abordagem Filtro		Técnicas de pesagem	
		Árvore de regressão [M5 Prime]	Algoritmo Greedy [estratégia forward e critério AIC]	Com novas variáveis	Com variáveis representativas dos conceitos	Heurística FSS [estratégia backward e critério CFS]	Heurística FSS [estratégia forward e critério CFS]	PCA *Weighting*	SVM *Weighting*
Modelos	**M0**	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
Capacidade explicativa	**0,545**	0,545	0,544	0,437	0,422	**0,499**	0,372	0,283	**0,433**
V01_AREA	**0,000**	0,000	0,000			**0,000**	0,000	0,000	**0,000**
V02_MACROZONA CENTRO AVR COD	**0,000**	0,000	0,000		0,000			0,000	
V03_MACROZONA CENTRO ILH COD	**0,007**	0,003	0,096						
V04_MACROZONA PRAIAS COD	**0,000**	0,000	0,000		0,000	**0,000**	0,000		**0,000**
V05_PRESERVACAO_NOVO	**0,000**	0,000	0,000						
V06_PRESERVACAO_USADO_CONSERVACAO&IDADE	**0,000**	0,000	0,000		0,000	**0,000**			**0,000**
V07_TIPOLOGIA_APART	**0,000**	0,000	0,000		0,000			0,000	
V08_TIPO_MORADIA	**0,000**	0,000	0,000					0,000	
V09_ATRIB_1_DUPLEX	**0,018**	0,019	0,017		0,000				
V10_ATRB_2_ARREC	**0,459**				0,068				
V11_ATRB_3_VARANDA	**0,002**	0,003	0,002						

V12_ATRB_4_SOTAO	**0,967**								
V13_ATRB_5_TERRACO	**0,000**	0,000	0,000						
V14_ATRB_6_LUGARGARAGEM	**0,168**								
V15_ATRB_7_GARAGEM	**0,000**	0,000	0,000		0,000				
V16_ATRB_8_PISO	**0,011**	0,012	0,012		0,147				
V17_ATRB_9_AQUECIMENTO	**0,000**	0,000	0,000						
V18_ATRB_10_WC	**0,000**	0,000	0,000						
V19_ATRB_11_REMODELADO	**0,000**	0,000	0,000						
V20_ATRB_12_LAREIRA	**0,000**	0,000	0,000			**0,000**			
V21_ATRB_13_HIDROMASSAGEM	**0,000**	0,000	0,000						
V22_POT_AMENIDADE_COMERCIO_T1	**0,000**	0,000	0,000		0,801	**0,000**			
V23_POT_AMENIDADE_COMERCIO_T2	**0,005**	0,000						0,000	
V24_POT_AMENIDADE_COMERCIO_T3	**0,009**	0,007	0,008						
V25_POT_AMENIDADE_CULTURA_T1	**0,005**	0,001	0,000					0,755	
V26_POT_AMENIDADE_CULTURA_T2	**0,751**				0,013				
V27_POT_AMENIDADE_CULTURA_T3	**0,000**	0,000	0,000						
V28_POT_AMENIDADE_DESPORTO_T1	**0,001**	0,001	0,001						
V29_POT_AMENIDADE_DESPORTO_T2	**0,011**	0,002	0,000						
V30_POT_AMENIDADE_DESPORTO_T3	**0,000**	0,000	0,000		0,000				
V31_POT_AMENIDADE_DIVERTIMENTOS_T2	**0,000**	0,000	0,000		0,001	**0,000**			
V32_POT_AMENIDADE_DIVERTIMENTOS_T3	**0,000**	0,000	0,000		0,001	**0,000**		0,000	
V33_POT_AMENIDADE_EDUCACAO_T1	**0,471**								
V35_POT_AMENIDADE_EDUCACAO_T3	**0,001**	0,000	0,000						
V36_POT_AMENIDADE_ELEMTURISTICOS_T1	**0,049**								
V37_POT_AMENIDADE_MOBILIDADE_T1	**0,000**	0,000	0,000						

V38_POT_AMENIDADE_MOBILIDADE_T2	**0,000**	0,000	0,000		0,882				
V39_POT_AMENIDADE_AREASVERDES_T1	**0,000**	0,000	0,000		0,000				
V40_POT_AMENIDADE_SERVSAUDE_T1	**0,237**								
V41_POT_AMENIDADE_SERVSAUDE_T2	**0,000**	0,000	0,000					0,000	
V42_POT_AMENIDADE_SERVSAUDE_T3	**0,000**	0,000				**0,000**		0,000	
V43_POT_AMENIDADE__ZI_T2	**0,000**	0,000	0,000						
V44_POT_AMENIDADE__ZI_T3	**0,000**	0,000	0,000					0,000	
V45_POT_AMENIDADE_UTILIDADES_T1	**0,000**	0,000	0,000						
V46_POT_AMENIDADE_UTILIDADES_T2	**0,000**	0,000	0,000		0,166				
V47_POT_AMENIDADE_UTILIDADES_T3	**0,000**	0,000						0,000	

VI. ANÁLISE FINAL E CONCLUSÃO

A identificação de atributos determinantes do preço da habitação é um objectivo elementar de todos os agentes que, de forma directa ou indirecta, estão envolvidos no estudo deste tema.

Na actividade de planeamento e ordenamento do território por exemplo, identificar o conjunto de indicadores chave na formação do preço da habitação permite oferecer à população soluções de ordenamento (e desenho urbano) que promovam os atributos que estas mais valorizam. A quantificação do valor destes atributos proporcionada pelas ferramentas econométrica permite, de forma complementar, fundamentar as diferentes opções de planeamento através, por exemplo, da avaliação custo – benefício. Os exemplos que podem ser referidos para justificar a importância deste trabalho são vários e não se esgotam no urbanismo.

Quando o estudo do mecanismo de formação de preços da habitação é essencial, ao investigador, com maior ou menor conhecimento teórico sobre cada um dos atributos da habitação, coloca-se o desafio da recolha de informação, a que necessariamente corresponde uma importante tarefa de selecção de qual o conjunto de atributos efectivamente relevante.

Inicialmente dependentes de informação produzida, de forma agregada, por instituições públicas e privadas, ou pelo desenvolvimento de mecanismos de recolha próprios – processos direccionados para conjuntos de atributos previamente identificados pelo investigador – os desafios de selecção encontravam-se camuflados pelas limitações do processo e dos dados efectivamente disponíveis. Com a maior disponibilidade de informação, torna-se evidente a dificuldade de o investigador ir muito mais além do que o conhecimento intrínseco que possui para seleccionar os correctos atributos, que lhe permitam identificar os determinantes do preço da habitação.

Actualmente, as possibilidades de recolha de informação tornaram as tarefas de análise de dados extremamente complexas. Como resposta científica, desenvolveram-se novas abordagens na análise de dados que, cruzando variadas disciplinas científicas,

fornece ferramentas expeditas para a manipulação de dados com níveis de complexidade muito elevados, permitindo a (semi) automatização de muitas tarefas. Os dois casos de estudo apresentados são exemplares neste aspecto: a utilização de algoritmos de selecção é um factor muito relevante para a construção de modelos eficazes que permitem identificar o conjunto restrito de atributos da habitação efectivamente relevante na formação do preço.

Autores como Caruama et al (1994) apontam várias vantagens para a utilização destes mecanismos de selecção (semi) automatizados os quais encontramos neste trabalho:

- É muitas vezes difícil determinar quais os efeitos que diferentes combinações de atributos têm no processo de construção de um modelo para determinado conjunto de dados. A selecção manual de atributos é complexa e leva, frequentemente, a uma selecção inferior.
- Fornece a liberdade ao investigador para identificar muito mais atributos potencialmente úteis, deixando o sistema determinar automaticamente quais usar.
- Permite adicionar facilmente novos atributos. O que é relevante para domínios em que a realidade muda muito rapidamente.

A reflexão teórica inicial permitiu aprofundar o estudo do tema habitação. As consequências dos fenómenos socioeconómicos e os padrões territoriais consequentes, permitem construir uma base teórica para escolher e recolher um conjunto de atributos que naturalmente tenham expectáveis relações com o preço da habitação.

No primeiro caso de estudo, enfrentamos um problema de menor complexidade, tanto no número de casos como de variáveis, como ainda dos conceitos subjacentes. O fenómeno é tratado de forma agregada, utilizando variáveis médias ou relativas, nas quais os possíveis efeitos específicos que criam dificuldades ao investigador são diluídos. No fundo, trata-se de uma análise na mesma escala da apresentada na reflexão teórica inicial, o que permite direccionar o trabalho a partir de conhecimento teórico adquirido.

Apesar de tudo, sujeitou-se o conjunto de variáveis recolhidas à utilização de técnicas de selecção. Conclui-se que apesar de um maior domínio do tema, estas técnicas são essenciais uma vez que permitem construir um modelo com melhor capacidade explicativa que o modelo base, inicial, sem qualquer técnica de selecção. Destaca-se a selecção de um conjunto muito restrito de variáveis – 5 – quando comparado com os 24 atributos recolhidos inicialmente.

As variáveis do melhor modelo construído referem-se a conceitos latentes, obtidos pela combinação de todas as variáveis da base de dados. O modelo aponta, como determinantes do preço, o "volume de construção" (o atributo mais importante), a "idade" e a "preservação" (aqui indicados pela sua ordem de importância). As restantes duas variáveis não mostraram uma relação linear significante com o valor médio dos prédios transaccionados nos municípios.

Concluindo, para o nível macro, comprovamos que o conhecimento e estudo dos fenómenos associados ao tema habitação, permite seleccionar conjuntos de variáveis determinantes do preço da habitação. No entanto, a recolha não é totalmente eficiente, sendo que a utilização de técnicas de selecção permite tornar ainda mais consistente e simples a construção de um modelo explicativo do valor médio dos prédios transaccionados nos municípios de Portugal continental.

No segundo caso de estudo, os resultados são mais cautelosos, embora apontem também para uma importante redução do número de variáveis que é efectivamente necessária para descrever o fenómeno.

No entanto, neste caso de estudo, salienta-se:

- a capacidade explicativa global dos modelos é muito mais reduzida (uma média de 45% para todos os modelos) muito embora se parta de um maior número de variáveis (e de casos).

- a redução de variáveis, obtida com as técnicas de selecção, conduz, invariavelmente, a uma menor capacidade explicativa, quando comparadas com o modelo base, que utiliza todo o conjunto de variáveis.

As explicações mais plausíveis para estes resultados podem ser associadas a:

- Verificação questionável dos pressupostos económicos assumidos no *modelo de preços hedónicos* (mercado único, bens razoavelmente semelhantes, concorrência perfeita, situação de equilíbrio).
- Errada formulação do modelo e da técnica de modelação seleccionada (modelo linear tradicional versus alternativas lineares e não lineares comuns na literatura).
- Não utilização de informação temporal associada. Os dados correspondem um longo período de 10 anos – de 2001 a 2010. Naturalmente este é um período extremamente longo para assumir consistentemente a inexistência de alterações temporais nas dinâmicas de mercado.
- Recolha insuficiente de atributos (que poderá ter excluído atributos importantes).

Apesar das limitações interpretativas descritas, destaca-se a importância de seleccionar correctamente os atributos necessários, uma vez que é possível reduzir em 83% o número de variáveis, perdendo apenas 8% de capacidade explicativa do preço por metro quadrado das habitações (para o melhor modelo construído).

O modelo que apresenta a melhor relação entre a capacidade explicativa e o número de variáveis seleccionadas, recorre a um mecanismo de filtro, para a selecção automatizada de atributos, que aplica uma heurística de busca utilizando um critério de selecção supervisionado dado pela medida CFS. Assim, é possível descrever o preço por metro quadrado da habitação nos territórios de Aveiro e Ílhavo, com um conjunto de 8 atributos os quais, tal como a reflexão teórica sugere, apontam para a importância dos atributos físicos e de localização, associados a cada habitação. Neste modelo, é naturalmente a área o atributo que maior variação induz no preço por metro quadrado de uma habitação – um efeito que traduz também economias de escala, visto que o sinal é negativo, indicando que quanto maior a área menor é o preço por metro quadrado. O segundo atributo mais relevante é o grau de preservação das habitações usadas. A

relação com o preço é naturalmente negativa: quanto pior for o estado de preservação, menor o valor por metro quadrado.

Ao nível da localização, as variáveis genéricas e de natureza classificatória mostraram a diferenciação das zonas balneares. A localização de uma habitação nas praias representa o terceiro atributo com maior impacto no preço da habitação (positivo). Este é um aspecto interessante e, mais uma vez, vem corroborar uma realidade identificada na reflexão teórica. As áreas balneares são um elemento importante sociologicamente para a maioria da população, a que acresce o facto de as praias em si, constituírem um recurso espacialmente limitado, localizado num dado local: consequentemente existe uma escassez natural dado o volume elevado da procura.

Ainda dentro da análise do modelo com melhor comportamento para o caso de estudo micro, identificam-se os atributos relativos às características de vizinhança. A ordem de importância com que surgem no modelo refere-se a: vizinhança aos serviços de saúde do Tipo3, vizinhança a divertimentos do Tipo2, vizinhança a divertimentos do Tipo3 e vizinhança de comércio do Tipo1. Se analisarmos as figuras 14 a 17 do anexo VIII.1 e a figura 10 da secção V.2.1, verificamos que a distribuição espacial dos atributos é consideravelmente coincidente com os atributos de localização, facto a que não é alheia a distribuição natural dos pontos de interesse, que se concentram efectivamente nas grandes centralidades atrás definidas. Embora se verifique sobreposição, os atributos de vizinhança não foram removidos, indicando que apresentam informação diferente da associada às grandes centralidades determinadas pelos atributos localização.

A sobreposição espacial entre atributos de localização e de vizinhança alerta para a possível imprecisão associada à construção das variáveis. Reafirma-se a necessidade de maior investigação.

À escala espacial da habitação, vista de forma individual, a complexidade de garantir todas as condições necessárias para a correcta aplicação de um *modelo de preços hedónicos* implica desdobrar o trabalho de investigação com o objectivo de aprofundar os diferentes pressupostos que foram sendo assumidos.

Apontam-se como os principais aspectos que devem ser melhorados numa investigação futura, as seguintes questões:

- Determinação da melhor formulação para a relação preço atributos. Muitos outros trabalhos científicos recorrem a formulações alternativas à simples relação linear. De facto, adaptações como a log-linear são importantes quando lidamos com atributos que não verificam efeitos constantes à escala ("mais, pode não significar, melhor"), dados pela formulação linear simples.

- Avaliar a componente temporal inerente aos dados. Como se salienta para a análise à micro escala, a utilização de dados de um período de 10 anos é um elemento indutor de potenciais erros de análise, visto que as realidades de mercado podem ter sofrido alterações muito significativas, tanto do lado da oferta como da procura (de notar que a reconhecida crise económica dos últimos anos tem exercido impactos significativos no mercado imobiliário nacional).

- Desenvolvimento e construção, fundamentada, dos novos atributos. É necessário um esforço que permita distinguir, de forma unívoca, os atributos de localização e de vizinhança. Ao mesmo tempo, é necessário rever toda a construção dos atributos de vizinhança aqui incorporados, os quais partiram de muitas considerações empíricas sem uma validação prévia

- Exploração da existência de submercados. Tipicamente, um submercado é considerado como o conjunto de habitações em que as suas características as tornam substitutos perfeitos (ou quase perfeitos). Esta definição tradicional enfrenta dificuldades tanto na identificação de quais as habitações que são efectivamente substitutos (quase) perfeitos e qual o consequente nível de agregação ou desagregação que deverá ser considerados. Diversos autores têm apontado múltiplas abordagens (consideração de divisões dadas por especialistas / comerciais, delimitações estatísticas) em múltiplas direcções (submercados espaciais, submercados tipológicos, submercados de idade / preservação). É comum, à esmagadora maioria destes estudos, apresentarem uma melhoria da capacidade explicativa dos modelos construídos.

- Testar a existência de interacções espaciais. A informação disponível sobre a habitação contém, cada vez mais dados com uma dimensão espacial. Tal

como refere Tse (2002), os efeitos de diferentes atributos tendem a variar com as diferentes localizações geográficas da habitação, o que resulta da heterogeneidade espacial. Bourassa et al (2007), argumenta também que é possível encontrar aglomerações de habitações com características similares, numa dada localização, que partilham amenidades comuns – ou seja, refere a possibilidade de existência de dependência espacial. Ambos os fenómenos espaciais são susceptíveis de afectar a estimação dada pela regressão linear, sendo que ambos os autores referem a necessidade de implementar ferramentas, baseadas na decomposição do factor estocástico, para a depuração dos modelos obtidos.

O trabalho permitiu ainda demonstrar as potencialidades das técnicas de *data mining*, suportadas por *software* intuitivo – como o utilizado neste trabalho – ao introduzir maior facilidade na tarefa de extracção de conhecimento de dados. A utilização deste *software*, gratuito e *open source*, revelou-se uma excelente alternativa à utilização do tradicional de *software* proprietário de análise de dados difundido nas ciências sociais (como é o caso especial do SPSS), o qual apresenta custos assinaláveis.

A interface clara e funcional, a rapidez e multiplicidade de operadores disponíveis, são pontos a favor da adopção do *RapidMiner*. Contudo, em certas tarefas tradicionais de análise de dados, a solução utilizada apresenta algumas limitações, das quais se identificaram neste trabalho: a menor capacidade de gerar, manipular e editar outputs, os menores recursos gráficos para análises prévias ou finais de vários tipos de variáveis, as opções, por vezes limitadas, de alguns operadores, para realizar procedimentos que são comuns nas metodologias tradicionais.

Por fim, é ainda de referir a dificuldade das soluções de *software* de análise de dados permitirem uma manipulação satisfatória de informação espacial, obrigando a que este tipo de análises continue a necessitar de *software* próprio para o seu tratamento.

VII. BIBLIOGRAFIA

ADAIR, ALASTAIR [et al] - House Prices and Accessibility: The Testing of Relationships within the Belfast Urban Area. Housing Studies. Vol. 15, Nº 5. 2000. p. 699-716.

ALBERGARIA, HENRIQUE [et al.] - A teoria da localização. Em: COSTA, J. ; NIJKAMP, P. - Compêndio de economia regional: teoria, temáticas e políticas. Princípia, 2009, p. 884.

AKAIKE, HIROTUGU - A new look at the statistical model identification. Automatic Control, IEEE Transactions.. Vol. 19, Nº6 1974, p.716-723.

BARANZINI, ANDREA [et al.] - Hedonic Methods in Housing Markets: Pricing Environmental Amenities and Segregation. Springer, 2008.

BATISTA, PAULO; [et al] - Actas: 16º Congresso da Associação Portuguesa de Desenvolvimento Regional, Preferências declaradas para a localização residencial. [em linha]. APDR – Associação Portuguesa de Desenvolvimento Regional. 2010. p. 3491.
[Consultado em. 01-08-2010]
[Disponível na internet em: <URL: http://www.apdr.pt/congresso/2010/PROCEEDINGS.html >].

BATISTA, PAULO; [et al.]. - Actas: XVII Jornadas de Classificação e Análise de Dados. *Técnicas de data mining na indução de modelos de formação de preço no mercado imobiliário em Portugal continental*. 2010

BOURASSA, STEVEN; [et al] - Spatial Dependence, Housing Submarkets, and House Price Prediction. The Journal of Real Estate Finance and Economics. Vol. 35, N.º 2. 2007. p. 143-160.

BOX, G; COX, D - An Analysis of transformations. Journal of the royal statistical society. Series B – Statistical methodology. Vol. 26, Nº 2. 1964. p. 211-252.

CARVALHO, JORGE - Ordenar a Cidade. Coimbra: Quarteto Editora, 2003. p.566

CHAPMAN, PETE [et al.]- Crisp-DM 1.0 - Step-by-step data mining guide. CRISP-DM consortium, 2000. 77p.
[Consultado em. 09-01-2010].
[Disponível na internet em:<URL:http://www.crisp-dm.org/CRISPWP-0800.pdf>]

CORREIA, PAULO - Políticas de solos no planeamento municipal. Fundação Calouste Gulbenkian, 2002. p. 403

FEYYAD, USAMA. - Data mining and knowledge discovery: making sense out of data. IEEE Expert. Vol. 11, N.º 5. 1996. p. 20-25.

GUERRA, ISABEL; [et al.] Plano Estratégico da Habitação [em linha]. IHRU - Instituto da Habitação e Reabilitação Urbana, 2008.
[Consultado em. 01-05-2010]
[Disponível na internet em: <URL:http://www.portaldahabitacao.pt/pt/ihru/estudos/ plano_estrategico/ documentos_plano_estrategico_habitacao.html>].

GUYON, ISABELLE [et al.] - Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning. Vol. 46, N.º 1. 2002. p. 389-422.

HAIR, JOSEPH [et al.] - Multivariate data analysis. 5th. New Jersey: Prentice-Hall International, 1998. 730 p.

HALL, MARK; SMITH, LLOYD; -. Proceedings: International Conference on Neural Information Processing and Intelligent Information Systems. Feature subset selection: a correlation based filter approach [em linha] Springer. 1997 p.855-858.
[Consultado em. 01-08-2010]
[Disponível em <URL:http://hdl.handle.net/10289/1515>].

INSTITUTO NACIONAL DE ESTATÍSTICA (INE) - Censos da População [em linha].
[Consultado em: 09-01-2010]
[Disponível na internet em: <URL: http://www.ine.pt/ >]

JUD, G. DONALD - The Effects of Zoning on Single-Family Residential Property Values: Charlotte, North Carolina. Land Economics. Vol. 56, Nº 2. 1980. p. 142-154.

KIEL, KATHERINE A.; ZABEL, JEFFREY E. - Location, location, location: The 3L Approach to house price determination. Journal of Housing Economics. Vol. 17, N.º 2. 2008. p. 175-190.

LI, MINGCHE.; BROWN, JAMES - Micro-Neighborhood Externalities and Hedonic Housing Prices. Land Economics. Vol. 56, N.º 2. 1980. p. 125-141.

LINNEMAN, PETER - Some empirical results on the nature of the hedonic price function for the urban housing market. Journal of Urban Economics. Vol. 8, N.º 1. 1980, p. 47-68.

LOPES, ANTÓNIO - Desenvolvimento regional: problemática, teoria, modelos. Fundação Calouste Gulbenkian. 2001. 406p.

LOPES, ANTÓNIO - O espaço económico. In: COSTA, J.; NIJKAMP, P.; - Compêndio de economia regional: teoria, temáticas e políticas. Princípia. 2009. 884 p..

MALPEZZI, STEPHEN - Hedonic Pricing Models: A Selective and Applied Review. Em: O'SULLIVAN, T.; GIBB, K.; - Housing Economics and Public Policy. Blackwell Science, 2008. 327 p.

MARQUES, JOÃO; [et al] - Actas do 16º Congresso da Associação Portuguesa de Desenvolvimento Regional, Funchal, 2010, O mercado habitacional - uma análise econométrica espacial. APDR – Associação Portuguesa de Desenvolvimento Regional. 2010. p.3491.
[Consultado em. 01-08-2010]
[Disponível na internet em: <URL: http://www.apdr.pt/congresso/2010/PROCEEDINGS.html >].

MARQUES, JOÃO; CASTRO, EDUARDO - Modelação do mercado da habitação. Em: VIEGAS, J. ; DENTINHO, T. - Desafios emergentes para o desenvolvimento regional. Princípia, 2010, 397 p..

MARTINS, BÁRBARA - O mercado da reabilitação [em linha]. Lisboa: AECOPS - Associação de Empresas de Construção, Obras Públicas e Serviços. 2009.
[Consultado em. 01-05-2010].
[Disponível na internet: <URL:http://www.aecops.pt/Home/MenuInstitucional/ Servi%C3% A7os/ Estudos/ tabid/132/language/pt-PT/Default.aspx>.

PALMQUIST, RAYMOND B. - Chapter 16 Property Value Models. In: KARL-GÖRAN, M.; JEFFREY, R. V. - Handbook of Environmental Economics. Elsevier. 2005. p. 763-819.

RAPID-I - RapidMiner 5.0 Manual [em linha]. Rapid-I GmbH, 2010. 169p.
[Consultado em. 01-08-2010]
[Disponível na internet em: <URL:http://sourceforge.net/projects/ rapidminer/ files/1.%20RapidMiner /5.0 /rapidminer- 5.0 -manual-english_v1.0.pdf/download>]

ROSS, JUSTIN; [et al] - Inconsistency in Welfare Inferences from Distance Variables in Hedonic Regressions. The Journal of Real Estate Finance and Economics. 2009. p.1-16.

SUCAHYOMONO, M.S. - Neighborhood impacts on suburban housing values. Dissertação para obtenção do grau de doutor. The Ohio State University. 2006. 223 p.

TAN, PANG-NING; [et al] - Introduction to data mining. Pearson Education. 2006. 769p.

TSE, RAYMOND Y. C. - Estimating Neighbourhood Effects in House Prices: Towards a New Hedonic Model Approach. Urban Studies. Vol. 39, N.º 7. 2002. p. 1165-1180.

VALENÇA, MÁRCIO Habitação no contexto da reestruturação económica. Análise Social. Vol. XXXVI, N.º 158 – 159, 2001 pp. 43-83.

WANG, Y.; WITTEN, I.; Induction of model trees for predicting continuous classes. Working paper 96/23. University of Waikato, Department of Computer Science. 1996.

WOOLDRIDGE, JEFFREY - Introductory Econometrics. 4ª Edição. Cengage Learning. 2009.

VIII. ANEXOS

VIII.1 – Atributos de vizinhança do melhor modelo hedónico micro

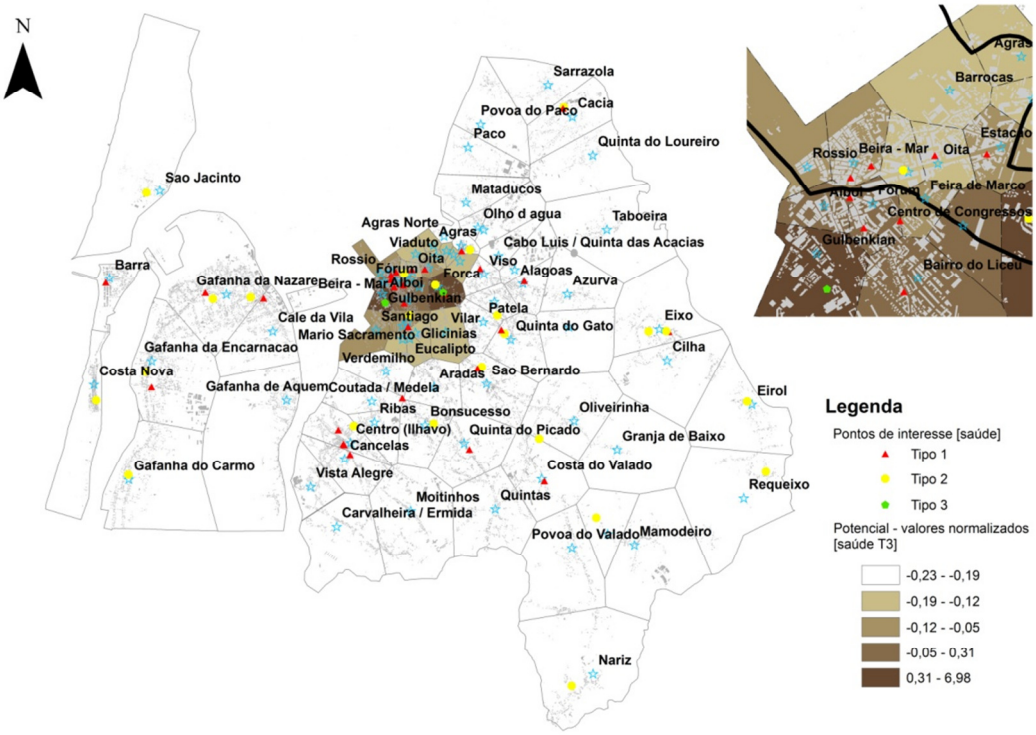

Figura 14 Representação dos valores normalizados de potencial, determinado pelos pontos de saúde do tipo 3.

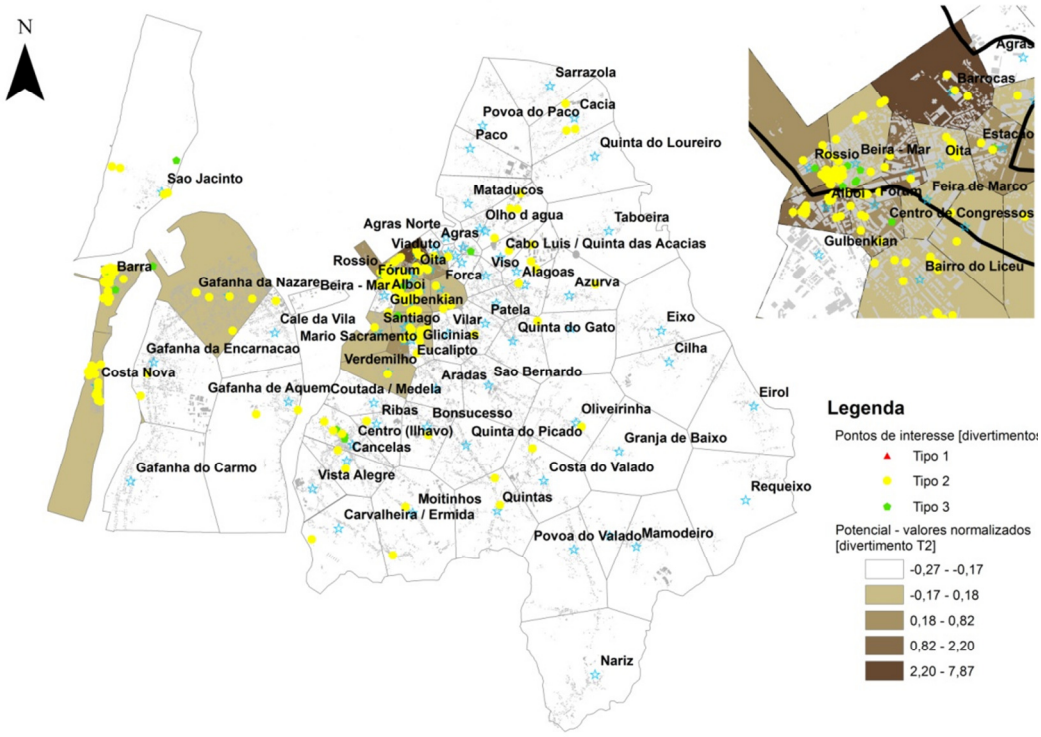

Figura 15 Representação dos valores normalizados de potencial, determinado pelos pontos de divertimentos do tipo 2.

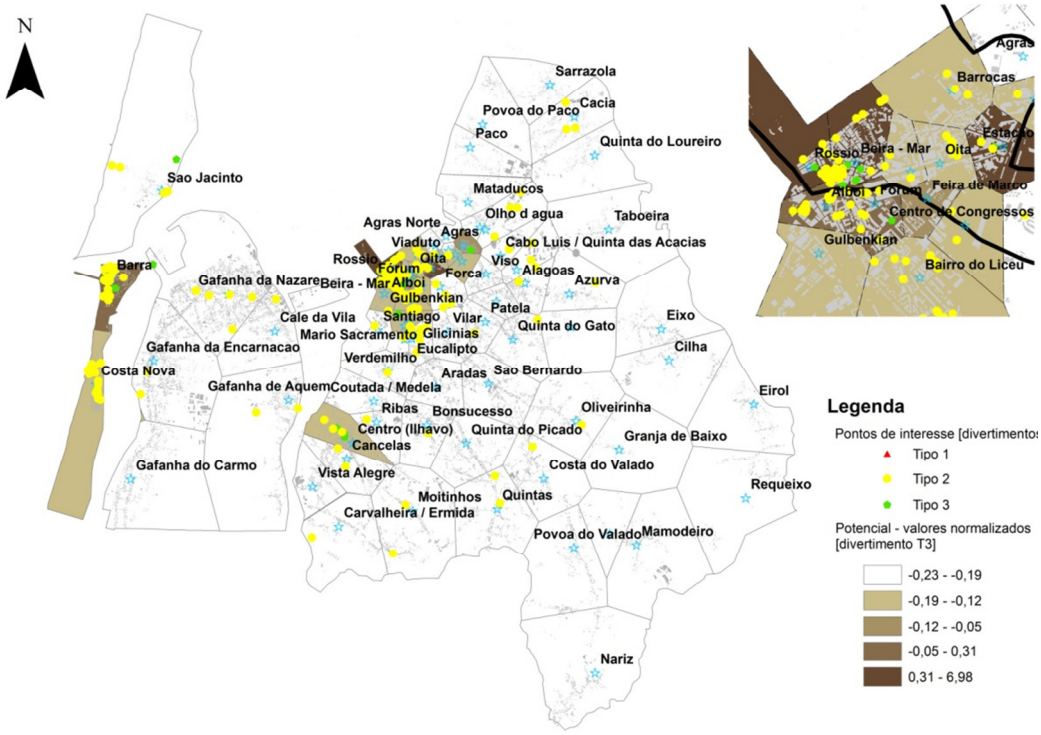

Figura 16 Representação dos valores normalizados de potencial, determinado pelos pontos de divertimento do tipo 3.

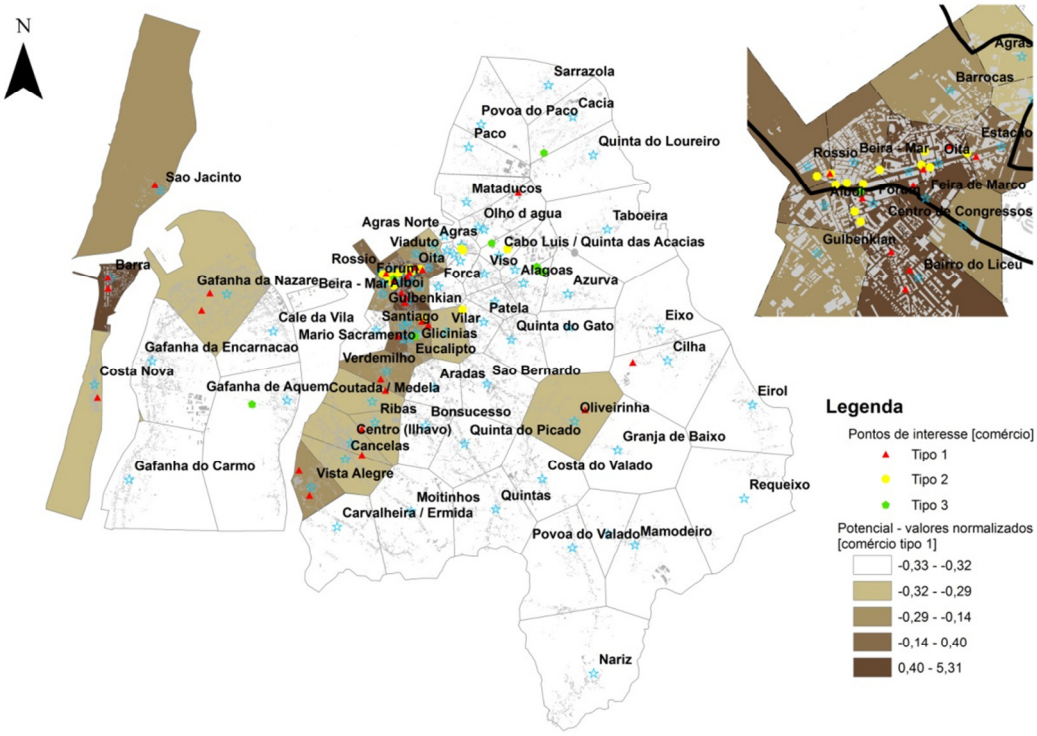

Figura 17 Representação dos valores normalizados de potencial, determinado pelos pontos de comércio do tipo 1.

VIII.2 – Código XML dos projectos implementados em RapidMiner

NOTA: Em ambas as escalas de análise (macro e micro), os projectos implementados em RapidMiner são semelhantes. Apenas existem duas diferenças significativas: o método de validação e, claro, as variáveis da base de dados.

Dadas as reduzidas diferenças, optou-se por apresentar o código xml dos projectos para análise dos dados à micro escala.

Modelos M0 M1 e M2


```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.0">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.0.11" expanded="true" name="Process">
    <process expanded="true" height="566" width="1083">
      <operator activated="true" class="retrieve" compatibility="5.0.11" expanded="true" height="60" name="Retrieve" width="90" x="47" y="35">
        <parameter key="repository_entry" value="BD 20101113"/>
      </operator>
      <operator activated="true" class="remove_duplicates" compatibility="5.0.11" expanded="true" height="76" name="Remove Duplicates (2)" width="90" x="45" y="120">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value="MUNICIPIO|PRECO|MICROZONA RECODE COD|MACROZONA|V01_AREA|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CONSERVACAO&amp;amp;IDADE|V05_PRESERVACAO_NOVO"/>
      </operator>
      <operator activated="true" class="work_on_subset" compatibility="5.0.11" expanded="false" height="76" name="Work on Subset" width="90" x="45" y="210">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value="V47_POT_AMENIDADE_UTILIDADES_T3|V46_POT_AMENIDADE_UTILIDADES_T2|V45_POT_AMENIDADE_UTILIDADES_T1|V44_POT_AMENIDADE__ZI_T3|V43_POT_AMENIDADE__ZI_T2|V42_POT_AMENIDADE_SERVSAUDE_T3|V41_POT_AMENIDADE_SERVSAUDE_T2|V40_POT_AMENIDADE_SERVSAUDE_T1|V39_POT_AMENIDADE_AREASVERDES_T1|V38_POT_AMENIDADE_MOBILIDADE_T2|V37_POT_AMENIDADE_MOBILIDADE_T1|V36_POT_AMENIDADE_ELEMTURISTICOS_T1|V35_POT_AMENIDADE_EDUCACAO_T3|V34_POT_AMENIDADE_EDUCACAO_T2|V33_POT_AMENIDADE_EDUCACAO_T1|V32_POT_AMENIDADE_DIVERTIMENTOS_T3|V31_POT_AMENIDADE_DIVERTIMENTOS_T2|V30_POT_AMENIDADE_DESPORTO_T3|V29_POT_AMENIDADE_DESPORTO_T2|V28_POT_AMENIDADE_DESPORTO_T1|V27_POT_AMENIDADE_CULTURA_T3|V26_POT_AMENIDADE_CULTURA_T2|V25_POT_AMENIDADE_CULTURA_T1|V24_POT_AMENIDADE_COMERCIO_T3|V23_POT_AMENIDADE_COMERCIO_T2|V22_POT_AMENIDADE_COMERCIO_T1|V21_ATRB_13_HIDROMASSAGEM|V20_ATRB_
```

12_LAREIRA|V19_ATRB_11_REMODELADO|V18_ATRB_10_WC|V17_ATRB_9_AQUECIMENTO|V16_ATRB_8_PISO|V15_ATRB_7_GARAGEM|V14_ATRB_6_LUGARGARAGEM|V13_ATRB_5_TERRACO|V12_ATRB_4_SOTAO|V11_ATRB_3_VARANDA|V10_ATRB_2_ARREC|V09_ATRIB_1_DUPLEX|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CONSERVACAO&IDADE|V05_PRESERVACAO_NOVO|V04_MACROZONA PRAIAS COD|V03_MACROZONA CENTRO ILH COD|V02_MACROZONA CENTRO AVR COD|V01_AREA|V00_PRECO_M2|MICROZONA RECODE COD|ID"/>
<parameter key="include_special_attributes" value="true"/>
<parameter key="keep_subset_only" value="true"/>
<process expanded="true" height="356" width="566">
<operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" height="76" name="Set Role" width="90" x="45" y="30">
<parameter key="name" value="ID"/>
<parameter key="target_role" value="id"/>
</operator>
<operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" height="76" name="Set Role (2)" width="90" x="179" y="30">
<parameter key="name" value="MICROZONA RECODE COD"/>
<parameter key="target_role" value="batch"/>
</operator>
<operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" height="76" name="Set Role (3)" width="90" x="313" y="30">
<parameter key="name" value="V00_PRECO_M2"/>
<parameter key="target_role" value="label"/>
</operator>
<connect from_port="exampleSet" to_op="Set Role" to_port="example set input"/>
<connect from_op="Set Role" from_port="example set output" to_op="Set Role (2)" to_port="example set input"/>
<connect from_op="Set Role (2)" from_port="example set output" to_op="Set Role (3)" to_port="example set input"/>
<connect from_op="Set Role (3)" from_port="example set output" to_port="example set"/>
<portSpacing port="source_exampleSet" spacing="0"/>
<portSpacing port="sink_example set" spacing="0"/>
<portSpacing port="sink_through 1" spacing="0"/>
</process>
</operator>
<operator activated="true" class="normalize" compatibility="5.0.11" expanded="true" height="94" name="Normalize" width="90" x="41" y="299">
<parameter key="attribute_filter_type" value="subset"/>
<parameter key="attributes" value="V47_POT_AMENIDADE_UTILIDADES_T3|V46_POT_AMENIDADE_UTILIDADES_T2|V45_POT_AMENIDADE_UTILIDADES_T1|V44_POT_AMENIDADE__ZI_T3|V43_POT_AMENIDADE__ZI_T2|V42_POT_AMENIDADE_SERVSAUDE_T3|V41_POT_AMENIDADE_SERVSAUDE_T2|V40_POT_AMENIDADE_SERVSAUDE_T1|V39_POT_AMENIDADE_AREASVERDES_T1|V38_POT_AMENIDADE_MOBILIDADE_T2|V37_POT_AMENIDADE_MOBILIDADE_T1|V36_POT_AMENIDADE_ELEMTURISTICOS_T1|V35_POT_AMENIDADE_EDUCACAO_T3|V34_POT_AMENIDADE_EDUCACAO_T2|V33_POT_AMENIDADE_EDUCACAO_T1|V32_POT_AMENIDADE_DIVERTIMENTOS_T3|V31_POT_AMENIDADE_DIVERTIMENTOS_T2|V30_POT_AMENIDADE_DESPORTO_T3|V29_POT_AMENIDADE_DESPORTO_T2|V28_POT_AMENIDADE_DESPORTO_T1|V27_POT_AMENIDADE_CULTURA_T3|V26_POT_AMENIDADE_CULTURA_T2|V25_POT_AMENIDADE_CULTURA_T1|V24_POT_AMENIDADE_COMERCIO_T3|V23_POT_AMENIDADE_COMERCIO_T2|V22_POT_AMENIDADE_COMERCIO_T1|V21_ATRB_13_HIDROMASSAGEM|V20_ATRB_12_LAREIRA|V19_ATRB_11_REMODELADO|V18_ATRB_10_WC|V17_ATRB_9_AQUECIMENTO|V16_ATRB_8_PISO|V15_ATRB_7_GARAGEM|V14_ATRB_6_LUGARGARAGEM|V13_ATRB_5_TERRACO|V12_ATRB_4_SOTAO|V11_ATRB_3_VARANDA|V10_ATRB_2_ARREC|V09_ATRIB_1_DUPLEX|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CONSERVACAO&IDADE|V05_PRESERVACAO_NOVO|V04_MACROZONA PRAIAS COD|V03_MACROZONA CENTRO ILH COD|V02_MACROZONA CENTRO AVR COD|V01_AREA"/>
</operator>

```
            <operator activated="true" class="remove_correlated_attributes" compatibility="5.0.11" expanded="true" height="76" name="Remove Correlated Attributes" width="90" x="45" y="435"/>
            <operator activated="true" class="split_data" compatibility="5.0.11" expanded="true" height="94" name="Split Data" width="90" x="246" y="165">
              <enumeration key="partitions">
                <parameter key="ratio" value="0.7"/>
                <parameter key="ratio" value="0.3"/>
              </enumeration>
              <parameter key="sampling_type" value="stratified sampling"/>
            </operator>
            <operator activated="true" class="multiply" compatibility="5.0.11" expanded="true" height="112" name="Multiply TREINO" width="90" x="380" y="75"/>
            <operator activated="true" class="linear_regression" compatibility="5.0.11" expanded="true" height="94" name="Linear Regression" width="90" x="648" y="30">
              <parameter key="feature_selection" value="none"/>
            </operator>
            <operator activated="true" class="multiply" compatibility="5.0.11" expanded="true" height="130" name="Multiply AVALIACAO" width="90" x="380" y="300"/>
            <operator activated="true" class="linear_regression" compatibility="5.0.11" expanded="true" height="94" name="Linear Regression [Greedy]" width="90" x="648" y="300">
              <parameter key="feature_selection" value="greedy"/>
            </operator>
            <operator activated="true" class="apply_model" compatibility="5.0.11" expanded="true" height="76" name="Apply Model Linear Regression [Greedy]" width="90" x="782" y="300">
              <list key="application_parameters"/>
            </operator>
            <operator activated="true" class="linear_regression" compatibility="5.0.11" expanded="true" height="94" name="Linear Regression [M5prime]" width="90" x="648" y="165"/>
            <operator activated="true" class="apply_model" compatibility="5.0.11" expanded="true" height="76" name="Apply Model Linear Regression [M5prime]" width="90" x="782" y="165">
              <list key="application_parameters"/>
            </operator>
            <operator activated="true" class="performance_regression" compatibility="5.0.11" expanded="true" height="76" name="Performance Linear Regression [M5prime]" width="90" x="916" y="165">
              <parameter key="root_mean_squared_error" value="false"/>
              <parameter key="squared_correlation" value="true"/>
            </operator>
            <operator activated="true" class="apply_model" compatibility="5.0.11" expanded="true" height="76" name="Apply Model Linear Regression [Sem FS]" width="90" x="782" y="30">
              <list key="application_parameters"/>
            </operator>
            <operator activated="true" class="performance_regression" compatibility="5.0.11" expanded="true" height="76" name="Performance Linear Regression [Sem FS]" width="90" x="916" y="30">
              <parameter key="root_mean_squared_error" value="false"/>
              <parameter key="squared_correlation" value="true"/>
            </operator>
            <operator activated="true" class="performance_regression" compatibility="5.0.11" expanded="true" height="76" name="Performance Linear Regression [Greedy]" width="90" x="916" y="300">
              <parameter key="root_mean_squared_error" value="false"/>
              <parameter key="squared_correlation" value="true"/>
            </operator>
            <connect from_op="Retrieve" from_port="output" to_op="Remove Duplicates (2)" to_port="example set input"/>
            <connect from_op="Remove Duplicates (2)" from_port="example set output" to_op="Work on Subset" to_port="example set"/>
            <connect from_op="Work on Subset" from_port="example set" to_op="Normalize" to_port="example set input"/>
```

```
            <connect from_op="Normalize" from_port="example set output" to_op="Remove Correlated Attributes" to_port="example set input"/>
            <connect from_op="Remove Correlated Attributes" from_port="example set output" to_op="Split Data" to_port="example set"/>
            <connect from_op="Split Data" from_port="partition 1" to_op="Multiply TREINO" to_port="input"/>
            <connect from_op="Split Data" from_port="partition 2" to_op="Multiply AVALIACAO" to_port="input"/>
            <connect from_op="Multiply TREINO" from_port="output 1" to_op="Linear Regression" to_port="training set"/>
            <connect from_op="Multiply TREINO" from_port="output 2" to_op="Linear Regression [M5prime]" to_port="training set"/>
            <connect from_op="Multiply TREINO" from_port="output 3" to_op="Linear Regression [Greedy]" to_port="training set"/>
            <connect from_op="Linear Regression" from_port="model" to_op="Apply Model Linear Regression [Sem FS]" to_port="model"/>
            <connect from_op="Multiply AVALIACAO" from_port="output 1" to_op="Apply Model Linear Regression [Sem FS]" to_port="unlabelled data"/>
            <connect from_op="Multiply AVALIACAO" from_port="output 2" to_op="Apply Model Linear Regression [M5prime]" to_port="unlabelled data"/>
            <connect from_op="Multiply AVALIACAO" from_port="output 3" to_op="Apply Model Linear Regression [Greedy]" to_port="unlabelled data"/>
            <connect from_op="Multiply AVALIACAO" from_port="output 4" to_port="result 7"/>
            <connect from_op="Linear Regression [Greedy]" from_port="model" to_op="Apply Model Linear Regression [Greedy]" to_port="model"/>
            <connect from_op="Apply Model Linear Regression [Greedy]" from_port="labelled data" to_op="Performance Linear Regression [Greedy]" to_port="labelled data"/>
            <connect from_op="Apply Model Linear Regression [Greedy]" from_port="model" to_port="result 6"/>
            <connect from_op="Linear Regression [M5prime]" from_port="model" to_op="Apply Model Linear Regression [M5prime]" to_port="model"/>
            <connect from_op="Apply Model Linear Regression [M5prime]" from_port="labelled data" to_op="Performance Linear Regression [M5prime]" to_port="labelled data"/>
            <connect from_op="Apply Model Linear Regression [M5prime]" from_port="model" to_port="result 4"/>
            <connect from_op="Performance Linear Regression [M5prime]" from_port="performance" to_port="result 3"/>
            <connect from_op="Apply Model Linear Regression [Sem FS]" from_port="labelled data" to_op="Performance Linear Regression [Sem FS]" to_port="labelled data"/>
            <connect from_op="Apply Model Linear Regression [Sem FS]" from_port="model" to_port="result 2"/>
            <connect from_op="Performance Linear Regression [Sem FS]" from_port="performance" to_port="result 1"/>
            <connect from_op="Performance Linear Regression [Greedy]" from_port="performance" to_port="result 5"/>
            <portSpacing port="source_input 1" spacing="0"/>
            <portSpacing port="sink_result 1" spacing="0"/>
            <portSpacing port="sink_result 2" spacing="0"/>
            <portSpacing port="sink_result 3" spacing="108"/>
            <portSpacing port="sink_result 4" spacing="0"/>
            <portSpacing port="sink_result 5" spacing="90"/>
            <portSpacing port="sink_result 6" spacing="0"/>
            <portSpacing port="sink_result 7" spacing="108"/>
            <portSpacing port="sink_result 8" spacing="0"/>
          </process>
        </operator>
</process>
```

Modelos M3

<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.0">
 <context>
  <input/>
  <output/>
  <macros/>
 </context>
 <operator activated="true" class="process" compatibility="5.0.11" expanded="true" name="Process">
  <process expanded="true" height="566" width="1036">
   <operator activated="true" class="retrieve" compatibility="5.0.11" expanded="true" height="60" name="Retrieve" width="90" x="45" y="30">
    <parameter key="repository_entry" value="BD 20101113"/>
   </operator>
   <operator activated="true" class="remove_duplicates" compatibility="5.0.11" expanded="true" height="76" name="Remove Duplicates (2)" width="90" x="45" y="120">
    <parameter key="attribute_filter_type" value="subset"/>
    <parameter key="attributes" value="MUNICIPIO|PRECO|MICROZONA RECODE COD|MACROZONA|V01_AREA|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CONSERVACAO&IDADE|V05_PRESERVACAO_NOVO"/>
   </operator>
   <operator activated="true" class="work_on_subset" compatibility="5.0.11" expanded="true" height="76" name="Work on Subset" width="90" x="45" y="210">
    <parameter key="attribute_filter_type" value="subset"/>
    <parameter key="attributes" value="V47_POT_AMENIDADE_UTILIDADES_T3|V46_POT_AMENIDADE_UTILIDADES_T2|V45_POT_AMENIDADE_UTILIDADES_T1|V44_POT_AMENIDADE__ZI_T3|V43_POT_AMENIDADE__ZI_T2|V42_POT_AMENIDADE_SERVSAUDE_T3|V41_POT_AMENIDADE_SERVSAUDE_T2|V40_POT_AMENIDADE_SERVSAUDE_T1|V39_POT_AMENIDADE_AREASVERDES_T1|V38_POT_AMENIDADE_MOBILIDADE_T2|V37_POT_AMENIDADE_MOBILIDADE_T1|V36_POT_AMENIDADE_ELEMTURISTICOS_T1|V35_POT_AMENIDADE_EDUCACAO_T3|V34_POT_AMENIDADE_EDUCACAO_T2|V33_POT_AMENIDADE_EDUCACAO_T1|V32_POT_AMENIDADE_DIVERTIMENTOS_T3|V31_POT_AMENIDADE_DIVERTIMENTOS_T2|V30_POT_AMENIDADE_DESPORTO_T3|V29_POT_AMENIDADE_DESPORTO_T2|V28_POT_AMENIDADE_DESPORTO_T1|V27_POT_AMENIDADE_CULTURA_T3|V26_POT_AMENIDADE_CULTURA_T2|V25_POT_AMENIDADE_CULTURA_T1|V24_POT_AMENIDADE_COMERCIO_T3|V23_POT_AMENIDADE_COMERCIO_T2|V22_POT_AMENIDADE_COMERCIO_T1|V21_ATRB_13_HIDROMASSAGEM|V20_ATRB_12_LAREIRA|V19_ATRB_11_REMODELADO|V18_ATRB_10_WC|V17_ATRB_9_AQUECIMENTO|V16_ATRB_8_PISO|V15_ATRB_7_GARAGEM|V14_ATRB_6_LUGARGARAGEM|V13_ATRB_5_TERRACO|V12_ATRB_4_SOTAO|V11_ATRB_3_VARANDA|V10_ATRB_2_ARREC|V09_ATRIB_1_DUPLEX|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CONSERVACAO&IDADE|V05_PRESERVACAO_NOVO|V04_MACROZONA PRAIAS COD|V03_MACROZONA CENTRO ILH COD|V02_MACROZONA CENTRO AVR COD|V01_AREA|V00_PRECO_M2|MICROZONA RECODE COD|ID"/>
    <parameter key="include_special_attributes" value="true"/>
    <parameter key="keep_subset_only" value="true"/>
    <process expanded="true">
     <operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" name="Set Role">
      <parameter key="name" value="ID"/>
      <parameter key="target_role" value="id"/>
     </operator>
     <operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" name="Set Role (2)">
      <parameter key="name" value="MICROZONA RECODE COD"/>
      <parameter key="target_role" value="batch"/>
     </operator>
     <operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" name="Set Role (3)">

```
        <parameter key="name" value="V00_PRECO_M2"/>
        <parameter key="target_role" value="label"/>
      </operator>
      <connect from_port="exampleSet" to_op="Set Role" to_port="example set input"/>
      <connect from_op="Set Role" from_port="example set output" to_op="Set Role (2)"
to_port="example set input"/>
      <connect from_op="Set Role (2)" from_port="example set output" to_op="Set Role (3)"
to_port="example set input"/>
      <connect from_op="Set Role (3)" from_port="example set output" to_port="example set"/>
      <portSpacing port="source_exampleSet" spacing="0"/>
      <portSpacing port="sink_example set" spacing="0"/>
      <portSpacing port="sink_through 1" spacing="0"/>
    </process>
  </operator>
  <operator activated="true" class="normalize" compatibility="5.0.11" expanded="true" height="94"
name="Normalize" width="90" x="45" y="300">
    <parameter key="attribute_filter_type" value="subset"/>
    <parameter key="attributes"
value="V47_POT_AMENIDADE_UTILIDADES_T3|V46_POT_AMENIDADE_UTILIDADES_T2|V4
5_POT_AMENIDADE_UTILIDADES_T1|V44_POT_AMENIDADE__ZI_T3|V43_POT_AMENIDAD
E__ZI_T2|V42_POT_AMENIDADE_SERVSAUDE_T3|V41_POT_AMENIDADE_SERVSAUDE_T2|
V40_POT_AMENIDADE_SERVSAUDE_T1|V39_POT_AMENIDADE_AREASVERDES_T1|V38_PO
T_AMENIDADE_MOBILIDADE_T2|V37_POT_AMENIDADE_MOBILIDADE_T1|V36_POT_AME
NIDADE_ELEMTURISTICOS_T1|V35_POT_AMENIDADE_EDUCACAO_T3|V34_POT_AMENIDA
DE_EDUCACAO_T2|V33_POT_AMENIDADE_EDUCACAO_T1|V32_POT_AMENIDADE_DIVER
TIMENTOS_T3|V31_POT_AMENIDADE_DIVERTIMENTOS_T2|V30_POT_AMENIDADE_DESPO
RTO_T3|V29_POT_AMENIDADE_DESPORTO_T2|V28_POT_AMENIDADE_DESPORTO_T1|V27_
POT_AMENIDADE_CULTURA_T3|V26_POT_AMENIDADE_CULTURA_T2|V25_POT_AMENIDA
DE_CULTURA_T1|V24_POT_AMENIDADE_COMERCIO_T3|V23_POT_AMENIDADE_COMERCI
O_T2|V22_POT_AMENIDADE_COMERCIO_T1|V21_ATRB_13_HIDROMASSAGEM|V20_ATRB_
12_LAREIRA|V19_ATRB_11_REMODELADO|V18_ATRB_10_WC|V17_ATRB_9_AQUECIMENT
O|V16_ATRB_8_PISO|V15_ATRB_7_GARAGEM|V14_ATRB_6_LUGARGARAGEM|V13_ATRB_5
_TERRACO|V12_ATRB_4_SOTAO|V11_ATRB_3_VARANDA|V10_ATRB_2_ARREC|V09_ATRIB
_1_DUPLEX|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CO
NSERVACAO&amp;amp;IDADE|V05_PRESERVACAO_NOVO|V04_MACROZONA PRAIAS
COD|V03_MACROZONA CENTRO ILH COD|V02_MACROZONA CENTRO AVR
COD|V01_AREA"/>
  </operator>
  <operator activated="true" class="remove_correlated_attributes" compatibility="5.0.11"
expanded="true" height="76" name="Remove Correlated Attributes" width="90" x="45" y="435"/>
  <operator activated="true" class="principal_component_analysis" compatibility="5.0.11"
expanded="true" height="94" name="PCA" width="90" x="246" y="165">
    <parameter key="dimensionality_reduction" value="fixed number"/>
    <parameter key="number_of_components" value="17"/>
  </operator>
  <operator activated="true" class="split_data" compatibility="5.0.11" expanded="true" height="94"
name="Split Data" width="90" x="380" y="300">
    <enumeration key="partitions">
      <parameter key="ratio" value="0.7"/>
      <parameter key="ratio" value="0.3"/>
    </enumeration>
    <parameter key="sampling_type" value="stratified sampling"/>
  </operator>
  <operator activated="true" class="linear_regression" compatibility="5.0.11" expanded="true"
height="94" name="Linear Regression" width="90" x="514" y="210">
    <parameter key="feature_selection" value="none"/>
  </operator>
  <operator activated="true" class="apply_model" compatibility="5.0.11" expanded="true"
height="76" name="Apply Model Linear Regression [Sem FS]" width="90" x="648" y="210">
    <list key="application_parameters"/>
```

```
    </operator>
    <operator activated="true" class="performance_regression" compatibility="5.0.11" expanded="true"
height="76" name="Performance Linear Regression [Sem FS]" width="90" x="782" y="210">
      <parameter key="root_mean_squared_error" value="false"/>
      <parameter key="squared_correlation" value="true"/>
    </operator>
    <connect from_op="Retrieve" from_port="output" to_op="Remove Duplicates (2)"
to_port="example set input"/>
    <connect from_op="Remove Duplicates (2)" from_port="example set output" to_op="Work on
Subset" to_port="example set"/>
    <connect from_op="Work on Subset" from_port="example set" to_op="Normalize"
to_port="example set input"/>
    <connect from_op="Normalize" from_port="example set output" to_op="Remove Correlated
Attributes" to_port="example set input"/>
    <connect from_op="Remove Correlated Attributes" from_port="example set output" to_op="PCA"
to_port="example set input"/>
    <connect from_op="PCA" from_port="example set output" to_op="Split Data" to_port="example
set"/>
    <connect from_op="PCA" from_port="preprocessing model" to_port="result 3"/>
    <connect from_op="Split Data" from_port="partition 1" to_op="Linear Regression"
to_port="training set"/>
    <connect from_op="Split Data" from_port="partition 2" to_op="Apply Model Linear Regression
[Sem FS]" to_port="unlabelled data"/>
    <connect from_op="Linear Regression" from_port="model" to_op="Apply Model Linear Regression
[Sem FS]" to_port="model"/>
    <connect from_op="Apply Model Linear Regression [Sem FS]" from_port="labelled data"
to_op="Performance Linear Regression [Sem FS]" to_port="labelled data"/>
    <connect from_op="Apply Model Linear Regression [Sem FS]" from_port="model" to_port="result
2"/>
    <connect from_op="Performance Linear Regression [Sem FS]" from_port="performance"
to_port="result 1"/>
    <portSpacing port="source_input 1" spacing="0"/>
    <portSpacing port="sink_result 1" spacing="288"/>
    <portSpacing port="sink_result 2" spacing="0"/>
    <portSpacing port="sink_result 3" spacing="0"/>
    <portSpacing port="sink_result 4" spacing="0"/>
  </process>
 </operator>
</process>
```

Modelos M4

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.0">
 <context>
  <input/>
  <output/>
  <macros/>
 </context>
 <operator activated="true" class="process" compatibility="5.0.11" expanded="true" name="Process">
  <process expanded="true" height="1929" width="2344">
   <operator activated="true" class="retrieve" compatibility="5.0.11" expanded="true" height="60"
name="Retrieve" width="90" x="45" y="30">
     <parameter key="repository_entry" value="BD 20101113"/>
   </operator>
```

```
    <operator activated="true" class="remove_duplicates" compatibility="5.0.11" expanded="true"
height="76" name="Remove Duplicates (2)" width="90" x="45" y="120">
      <parameter key="attribute_filter_type" value="subset"/>
      <parameter key="attributes" value="MUNICIPIO|PRECO|MICROZONA RECODE
COD|MACROZONA|V01_AREA|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERV
ACAO_USADO_CONSERVACAO&amp;amp;IDADE|V05_PRESERVACAO_NOVO"/>
    </operator>
    <operator activated="true" class="work_on_subset" compatibility="5.0.11" expanded="false"
height="76" name="Work on Subset" width="90" x="45" y="210">
      <parameter key="attribute_filter_type" value="subset"/>
      <parameter key="attributes"
value="V47_POT_AMENIDADE_UTILIDADES_T3|V46_POT_AMENIDADE_UTILIDADES_T2|V4
5_POT_AMENIDADE_UTILIDADES_T1|V44_POT_AMENIDADE__ZI_T3|V43_POT_AMENIDAD
E__ZI_T2|V42_POT_AMENIDADE_SERVSAUDE_T3|V41_POT_AMENIDADE_SERVSAUDE_T2|
V40_POT_AMENIDADE_SERVSAUDE_T1|V39_POT_AMENIDADE_AREASVERDES_T1|V38_PO
T_AMENIDADE_MOBILIDADE_T2|V37_POT_AMENIDADE_MOBILIDADE_T1|V36_POT_AME
NIDADE_ELEMTURISTICOS_T1|V35_POT_AMENIDADE_EDUCACAO_T3|V34_POT_AMENIDA
DE_EDUCACAO_T2|V33_POT_AMENIDADE_EDUCACAO_T1|V32_POT_AMENIDADE_DIVER
TIMENTOS_T3|V31_POT_AMENIDADE_DIVERTIMENTOS_T2|V30_POT_AMENIDADE_DESPO
RTO_T3|V29_POT_AMENIDADE_DESPORTO_T2|V28_POT_AMENIDADE_DESPORTO_T1|V27_
POT_AMENIDADE_CULTURA_T3|V26_POT_AMENIDADE_CULTURA_T2|V25_POT_AMENIDA
DE_CULTURA_T1|V24_POT_AMENIDADE_COMERCIO_T3|V23_POT_AMENIDADE_COMERCI
O_T2|V22_POT_AMENIDADE_COMERCIO_T1|V21_ATRB_13_HIDROMASSAGEM|V20_ATRB_
12_LAREIRA|V19_ATRB_11_REMODELADO|V18_ATRB_10_WC|V17_ATRB_9_AQUECIMENT
O|V16_ATRB_8_PISO|V15_ATRB_7_GARAGEM|V14_ATRB_6_LUGARGARAGEM|V13_ATRB_5
_TERRACO|V12_ATRB_4_SOTAO|V11_ATRB_3_VARANDA|V10_ATRB_2_ARREC|V09_ATRIB
_1_DUPLEX|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CO
NSERVACAO&amp;amp;IDADE|V05_PRESERVACAO_NOVO|V04_MACROZONA PRAIAS
COD|V03_MACROZONA CENTRO ILH COD|V02_MACROZONA CENTRO AVR
COD|V01_AREA|V00_PRECO_M2|MICROZONA RECODE COD|ID"/>
      <parameter key="include_special_attributes" value="true"/>
      <parameter key="keep_subset_only" value="true"/>
      <process expanded="true">
        <operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" name="Set
Role">
          <parameter key="name" value="ID"/>
          <parameter key="target_role" value="id"/>
        </operator>
        <operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" name="Set
Role (2)">
          <parameter key="name" value="MICROZONA RECODE COD"/>
          <parameter key="target_role" value="batch"/>
        </operator>
        <operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" name="Set
Role (3)">
          <parameter key="name" value="V00_PRECO_M2"/>
          <parameter key="target_role" value="label"/>
        </operator>
        <connect from_port="exampleSet" to_op="Set Role" to_port="example set input"/>
        <connect from_op="Set Role" from_port="example set output" to_op="Set Role (2)"
to_port="example set input"/>
        <connect from_op="Set Role (2)" from_port="example set output" to_op="Set Role (3)"
to_port="example set input"/>
        <connect from_op="Set Role (3)" from_port="example set output" to_port="example set"/>
        <portSpacing port="source_exampleSet" spacing="0"/>
        <portSpacing port="sink_example set" spacing="0"/>
        <portSpacing port="sink_through 1" spacing="0"/>
      </process>
    </operator>
```

```
    <operator activated="true" class="normalize" compatibility="5.0.11" expanded="true" height="94"
name="Normalize" width="90" x="45" y="300">
      <parameter key="attribute_filter_type" value="subset"/>
      <parameter key="attributes"
value="V47_POT_AMENIDADE_UTILIDADES_T3|V46_POT_AMENIDADE_UTILIDADES_T2|V4
5_POT_AMENIDADE_UTILIDADES_T1|V44_POT_AMENIDADE__ZI_T3|V43_POT_AMENIDAD
E__ZI_T2|V42_POT_AMENIDADE_SERVSAUDE_T3|V41_POT_AMENIDADE_SERVSAUDE_T2|
V40_POT_AMENIDADE_SERVSAUDE_T1|V39_POT_AMENIDADE_AREASVERDES_T1|V38_PO
T_AMENIDADE_MOBILIDADE_T2|V37_POT_AMENIDADE_MOBILIDADE_T1|V36_POT_AME
NIDADE_ELEMTURISTICOS_T1|V35_POT_AMENIDADE_EDUCACAO_T3|V34_POT_AMENIDA
DE_EDUCACAO_T2|V33_POT_AMENIDADE_EDUCACAO_T1|V32_POT_AMENIDADE_DIVER
TIMENTOS_T3|V31_POT_AMENIDADE_DIVERTIMENTOS_T2|V30_POT_AMENIDADE_DESPO
RTO_T3|V29_POT_AMENIDADE_DESPORTO_T2|V28_POT_AMENIDADE_DESPORTO_T1|V27_
POT_AMENIDADE_CULTURA_T3|V26_POT_AMENIDADE_CULTURA_T2|V25_POT_AMENIDA
DE_CULTURA_T1|V24_POT_AMENIDADE_COMERCIO_T3|V23_POT_AMENIDADE_COMERCI
O_T2|V22_POT_AMENIDADE_COMERCIO_T1|V21_ATRB_13_HIDROMASSAGEM|V20_ATRB_
12_LAREIRA|V19_ATRB_11_REMODELADO|V18_ATRB_10_WC|V17_ATRB_9_AQUECIMENT
O|V16_ATRB_8_PISO|V15_ATRB_7_GARAGEM|V14_ATRB_6_LUGARGARAGEM|V13_ATRB_5
_TERRACO|V12_ATRB_4_SOTAO|V11_ATRB_3_VARANDA|V10_ATRB_2_ARREC|V09_ATRIB
_1_DUPLEX|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CO
NSERVACAO&amp;amp;IDADE|V05_PRESERVACAO_NOVO|V04_MACROZONA PRAIAS
COD|V03_MACROZONA CENTRO ILH COD|V02_MACROZONA CENTRO AVR
COD|V01_AREA"/>
    </operator>
    <operator activated="true" class="remove_correlated_attributes" compatibility="5.0.11"
expanded="true" height="76" name="Remove Correlated Attributes" width="90" x="45" y="435"/>
    <operator activated="true" class="multiply" compatibility="5.0.11" expanded="true" height="364"
name="Multiply" width="90" x="447" y="975"/>
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA (17)" width="90" x="648" y="1515">
      <parameter key="component_number" value="17"/>
    </operator>
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA (13)" width="90" x="648" y="1425">
      <parameter key="component_number" value="16"/>
    </operator>
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA (16)" width="90" x="648" y="1335">
      <parameter key="component_number" value="15"/>
    </operator>
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA (12)" width="90" x="648" y="1245">
      <parameter key="component_number" value="14"/>
    </operator>
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA (11)" width="90" x="648" y="1155">
      <parameter key="component_number" value="13"/>
    </operator>
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA (10)" width="90" x="648" y="1065">
      <parameter key="component_number" value="12"/>
    </operator>
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA (15)" width="90" x="648" y="975">
      <parameter key="component_number" value="11"/>
    </operator>
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA (14)" width="90" x="648" y="885">
      <parameter key="component_number" value="10"/>
    </operator>
```

```
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA (5)" width="90" x="648" y="795">
      <parameter key="component_number" value="9"/>
    </operator>
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA (9)" width="90" x="648" y="660">
      <parameter key="component_number" value="8"/>
    </operator>
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights (12)" width="90" x="782" y="660">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA (6)" width="90" x="648" y="570">
      <parameter key="component_number" value="7"/>
    </operator>
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights (11)" width="90" x="782" y="570">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
    <operator activated="true" class="join" compatibility="5.0.11" expanded="true" height="76"
name="Join (8)" width="90" x="916" y="615"/>
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA (7)" width="90" x="648" y="480">
      <parameter key="component_number" value="6"/>
    </operator>
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights (10)" width="90" x="782" y="480">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA (8)" width="90" x="648" y="390">
      <parameter key="component_number" value="5"/>
    </operator>
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights (9)" width="90" x="782" y="390">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
    <operator activated="true" class="join" compatibility="5.0.11" expanded="true" height="76"
name="Join (7)" width="90" x="916" y="435"/>
    <operator activated="true" class="join" compatibility="5.0.11" expanded="true" height="76"
name="Join (9)" width="90" x="1050" y="525"/>
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA (4)" width="90" x="648" y="300">
      <parameter key="component_number" value="4"/>
    </operator>
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights (7)" width="90" x="782" y="300">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
```

```
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA (3)" width="90" x="648" y="210">
      <parameter key="component_number" value="3"/>
    </operator>
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights (5)" width="90" x="782" y="210">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
    <operator activated="true" class="join" compatibility="5.0.11" expanded="true" height="76"
name="Join (3)" width="90" x="916" y="255"/>
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA (2)" width="90" x="648" y="120">
      <parameter key="component_number" value="2"/>
    </operator>
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights (3)" width="90" x="782" y="120">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
    <operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true"
height="76" name="Weight by PCA" width="90" x="648" y="30"/>
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights" width="90" x="782" y="30">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
    <operator activated="true" class="join" compatibility="5.0.11" expanded="true" height="76"
name="Join" width="90" x="916" y="75"/>
    <operator activated="true" class="join" compatibility="5.0.11" expanded="true" height="76"
name="Join (4)" width="90" x="1050" y="165"/>
    <operator activated="true" class="join" compatibility="5.0.11" expanded="true" height="76"
name="Join (10)" width="90" x="1184" y="345"/>
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights (2)" width="90" x="782" y="795">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights (4)" width="90" x="782" y="1065">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights (6)" width="90" x="782" y="1155">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights (8)" width="90" x="782" y="1335">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
```

```
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights (13)" width="90" x="782" y="1425">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
    <operator activated="true" class="join" compatibility="5.0.11" expanded="true" height="76"
name="Join (13)" width="90" x="916" y="1380"/>
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights (14)" width="90" x="782" y="885">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
    <operator activated="true" class="join" compatibility="5.0.11" expanded="true" height="76"
name="Join (11)" width="90" x="916" y="840"/>
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights (15)" width="90" x="782" y="975">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
    <operator activated="true" class="join" compatibility="5.0.11" expanded="true" height="76"
name="Join (12)" width="90" x="916" y="1020"/>
    <operator activated="true" class="join" compatibility="5.0.11" expanded="true" height="76"
name="Join (6)" width="90" x="1050" y="930"/>
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights (16)" width="90" x="782" y="1245">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
    <operator activated="true" class="join" compatibility="5.0.11" expanded="true" height="76"
name="Join (2)" width="90" x="916" y="1200"/>
    <operator activated="true" class="join" compatibility="5.0.11" expanded="true" height="76"
name="Join (5)" width="90" x="1050" y="1290"/>
    <operator activated="true" class="join" compatibility="5.0.11" expanded="true" height="76"
name="Join (14)" width="90" x="1184" y="1110"/>
    <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true"
height="94" name="Select by Weights (17)" width="90" x="782" y="1515">
      <parameter key="weight_relation" value="top k"/>
      <parameter key="weight" value="0.7"/>
      <parameter key="k" value="1"/>
    </operator>
    <operator activated="true" class="join" compatibility="5.0.11" expanded="true" height="76"
name="Join (15)" width="90" x="1385" y="1335"/>
    <operator activated="true" class="join" compatibility="5.0.11" expanded="true" height="76"
name="Join (16)" width="90" x="1586" y="480"/>
    <operator activated="true" class="split_data" compatibility="5.0.11" expanded="true" height="94"
name="Split Data" width="90" x="1775" y="345">
      <enumeration key="partitions">
        <parameter key="ratio" value="0.7"/>
        <parameter key="ratio" value="0.3"/>
      </enumeration>
      <parameter key="sampling_type" value="stratified sampling"/>
    </operator>
    <operator activated="true" class="linear_regression" compatibility="5.0.11" expanded="true"
height="94" name="Linear Regression [Sem FS]" width="90" x="1976" y="345">
      <parameter key="feature_selection" value="none"/>
    </operator>
```

```
    <operator activated="true" class="apply_model" compatibility="5.0.11" expanded="true"
height="76" name="Apply Model Linear Regression [Sem FS]" width="90" x="2110" y="345">
      <list key="application_parameters"/>
    </operator>
    <operator activated="true" class="performance_regression" compatibility="5.0.11" expanded="true"
height="76" name="Performance Linear Regression [Sem FS]" width="90" x="2244" y="345">
      <parameter key="root_mean_squared_error" value="false"/>
      <parameter key="squared_correlation" value="true"/>
    </operator>
    <connect from_op="Retrieve" from_port="output" to_op="Remove Duplicates (2)"
to_port="example set input"/>
    <connect from_op="Remove Duplicates (2)" from_port="example set output" to_op="Work on
Subset" to_port="example set"/>
    <connect from_op="Work on Subset" from_port="example set" to_op="Normalize"
to_port="example set input"/>
    <connect from_op="Normalize" from_port="example set output" to_op="Remove Correlated
Attributes" to_port="example set input"/>
    <connect from_op="Remove Correlated Attributes" from_port="example set output"
to_op="Multiply" to_port="input"/>
    <connect from_op="Multiply" from_port="output 1" to_op="Weight by PCA" to_port="example
set"/>
    <connect from_op="Multiply" from_port="output 2" to_op="Weight by PCA (2)" to_port="example
set"/>
    <connect from_op="Multiply" from_port="output 3" to_op="Weight by PCA (3)" to_port="example
set"/>
    <connect from_op="Multiply" from_port="output 4" to_op="Weight by PCA (4)" to_port="example
set"/>
    <connect from_op="Multiply" from_port="output 5" to_op="Weight by PCA (8)" to_port="example
set"/>
    <connect from_op="Multiply" from_port="output 6" to_op="Weight by PCA (7)" to_port="example
set"/>
    <connect from_op="Multiply" from_port="output 7" to_op="Weight by PCA (6)" to_port="example
set"/>
    <connect from_op="Multiply" from_port="output 8" to_op="Weight by PCA (9)" to_port="example
set"/>
    <connect from_op="Multiply" from_port="output 9" to_op="Weight by PCA (5)" to_port="example
set"/>
    <connect from_op="Multiply" from_port="output 10" to_op="Weight by PCA (14)"
to_port="example set"/>
    <connect from_op="Multiply" from_port="output 11" to_op="Weight by PCA (15)"
to_port="example set"/>
    <connect from_op="Multiply" from_port="output 12" to_op="Weight by PCA (10)"
to_port="example set"/>
    <connect from_op="Multiply" from_port="output 13" to_op="Weight by PCA (11)"
to_port="example set"/>
    <connect from_op="Multiply" from_port="output 14" to_op="Weight by PCA (12)"
to_port="example set"/>
    <connect from_op="Multiply" from_port="output 15" to_op="Weight by PCA (16)"
to_port="example set"/>
    <connect from_op="Multiply" from_port="output 16" to_op="Weight by PCA (13)"
to_port="example set"/>
    <connect from_op="Multiply" from_port="output 17" to_op="Weight by PCA (17)"
to_port="example set"/>
    <connect from_op="Weight by PCA (17)" from_port="weights" to_op="Select by Weights (17)"
to_port="weights"/>
    <connect from_op="Weight by PCA (17)" from_port="example set" to_op="Select by Weights (17)"
to_port="example set input"/>
    <connect from_op="Weight by PCA (13)" from_port="weights" to_op="Select by Weights (13)"
to_port="weights"/>
```

```
    <connect from_op="Weight by PCA (13)" from_port="example set" to_op="Select by Weights (13)"
to_port="example set input"/>
    <connect from_op="Weight by PCA (16)" from_port="weights" to_op="Select by Weights (8)"
to_port="weights"/>
    <connect from_op="Weight by PCA (16)" from_port="example set" to_op="Select by Weights (8)"
to_port="example set input"/>
    <connect from_op="Weight by PCA (12)" from_port="weights" to_op="Select by Weights (16)"
to_port="weights"/>
    <connect from_op="Weight by PCA (12)" from_port="example set" to_op="Select by Weights (16)"
to_port="example set input"/>
    <connect from_op="Weight by PCA (11)" from_port="weights" to_op="Select by Weights (6)"
to_port="weights"/>
    <connect from_op="Weight by PCA (11)" from_port="example set" to_op="Select by Weights (6)"
to_port="example set input"/>
    <connect from_op="Weight by PCA (10)" from_port="weights" to_op="Select by Weights (4)"
to_port="weights"/>
    <connect from_op="Weight by PCA (10)" from_port="example set" to_op="Select by Weights (4)"
to_port="example set input"/>
    <connect from_op="Weight by PCA (15)" from_port="weights" to_op="Select by Weights (15)"
to_port="weights"/>
    <connect from_op="Weight by PCA (15)" from_port="example set" to_op="Select by Weights (15)"
to_port="example set input"/>
    <connect from_op="Weight by PCA (14)" from_port="weights" to_op="Select by Weights (14)"
to_port="weights"/>
    <connect from_op="Weight by PCA (14)" from_port="example set" to_op="Select by Weights (14)"
to_port="example set input"/>
    <connect from_op="Weight by PCA (5)" from_port="weights" to_op="Select by Weights (2)"
to_port="weights"/>
    <connect from_op="Weight by PCA (5)" from_port="example set" to_op="Select by Weights (2)"
to_port="example set input"/>
    <connect from_op="Weight by PCA (9)" from_port="weights" to_op="Select by Weights (12)"
to_port="weights"/>
    <connect from_op="Weight by PCA (9)" from_port="example set" to_op="Select by Weights (12)"
to_port="example set input"/>
    <connect from_op="Select by Weights (12)" from_port="example set output" to_op="Join (8)"
to_port="right"/>
    <connect from_op="Weight by PCA (6)" from_port="weights" to_op="Select by Weights (11)"
to_port="weights"/>
    <connect from_op="Weight by PCA (6)" from_port="example set" to_op="Select by Weights (11)"
to_port="example set input"/>
    <connect from_op="Select by Weights (11)" from_port="example set output" to_op="Join (8)"
to_port="left"/>
    <connect from_op="Join (8)" from_port="join" to_op="Join (9)" to_port="right"/>
    <connect from_op="Weight by PCA (7)" from_port="weights" to_op="Select by Weights (10)"
to_port="weights"/>
    <connect from_op="Weight by PCA (7)" from_port="example set" to_op="Select by Weights (10)"
to_port="example set input"/>
    <connect from_op="Select by Weights (10)" from_port="example set output" to_op="Join (7)"
to_port="right"/>
    <connect from_op="Weight by PCA (8)" from_port="weights" to_op="Select by Weights (9)"
to_port="weights"/>
    <connect from_op="Weight by PCA (8)" from_port="example set" to_op="Select by Weights (9)"
to_port="example set input"/>
    <connect from_op="Select by Weights (9)" from_port="example set output" to_op="Join (7)"
to_port="left"/>
    <connect from_op="Join (7)" from_port="join" to_op="Join (9)" to_port="left"/>
    <connect from_op="Join (9)" from_port="join" to_op="Join (10)" to_port="right"/>
    <connect from_op="Weight by PCA (4)" from_port="weights" to_op="Select by Weights (7)"
to_port="weights"/>
```

<connect from_op="Weight by PCA (4)" from_port="example set" to_op="Select by Weights (7)" to_port="example set input"/>
<connect from_op="Select by Weights (7)" from_port="example set output" to_op="Join (3)" to_port="right"/>
<connect from_op="Weight by PCA (3)" from_port="weights" to_op="Select by Weights (5)" to_port="weights"/>
<connect from_op="Weight by PCA (3)" from_port="example set" to_op="Select by Weights (5)" to_port="example set input"/>
<connect from_op="Select by Weights (5)" from_port="example set output" to_op="Join (3)" to_port="left"/>
<connect from_op="Join (3)" from_port="join" to_op="Join (4)" to_port="right"/>
<connect from_op="Weight by PCA (2)" from_port="weights" to_op="Select by Weights (3)" to_port="weights"/>
<connect from_op="Weight by PCA (2)" from_port="example set" to_op="Select by Weights (3)" to_port="example set input"/>
<connect from_op="Select by Weights (3)" from_port="example set output" to_op="Join" to_port="right"/>
<connect from_op="Weight by PCA" from_port="weights" to_op="Select by Weights" to_port="weights"/>
<connect from_op="Weight by PCA" from_port="example set" to_op="Select by Weights" to_port="example set input"/>
<connect from_op="Select by Weights" from_port="example set output" to_op="Join" to_port="left"/>
<connect from_op="Join" from_port="join" to_op="Join (4)" to_port="left"/>
<connect from_op="Join (4)" from_port="join" to_op="Join (10)" to_port="left"/>
<connect from_op="Join (10)" from_port="join" to_op="Join (16)" to_port="left"/>
<connect from_op="Select by Weights (2)" from_port="example set output" to_op="Join (11)" to_port="left"/>
<connect from_op="Select by Weights (4)" from_port="example set output" to_op="Join (12)" to_port="right"/>
<connect from_op="Select by Weights (6)" from_port="example set output" to_op="Join (2)" to_port="left"/>
<connect from_op="Select by Weights (8)" from_port="example set output" to_op="Join (13)" to_port="left"/>
<connect from_op="Select by Weights (13)" from_port="example set output" to_op="Join (13)" to_port="right"/>
<connect from_op="Join (13)" from_port="join" to_op="Join (5)" to_port="right"/>
<connect from_op="Select by Weights (14)" from_port="example set output" to_op="Join (11)" to_port="right"/>
<connect from_op="Join (11)" from_port="join" to_op="Join (6)" to_port="left"/>
<connect from_op="Select by Weights (15)" from_port="example set output" to_op="Join (12)" to_port="left"/>
<connect from_op="Join (12)" from_port="join" to_op="Join (6)" to_port="right"/>
<connect from_op="Join (6)" from_port="join" to_op="Join (14)" to_port="left"/>
<connect from_op="Select by Weights (16)" from_port="example set output" to_op="Join (2)" to_port="right"/>
<connect from_op="Join (2)" from_port="join" to_op="Join (5)" to_port="left"/>
<connect from_op="Join (5)" from_port="join" to_op="Join (14)" to_port="right"/>
<connect from_op="Join (14)" from_port="join" to_op="Join (15)" to_port="left"/>
<connect from_op="Select by Weights (17)" from_port="example set output" to_op="Join (15)" to_port="right"/>
<connect from_op="Join (15)" from_port="join" to_op="Join (16)" to_port="right"/>
<connect from_op="Join (16)" from_port="join" to_op="Split Data" to_port="example set"/>
<connect from_op="Split Data" from_port="partition 1" to_op="Linear Regression [Sem FS]" to_port="training set"/>
<connect from_op="Split Data" from_port="partition 2" to_op="Apply Model Linear Regression [Sem FS]" to_port="unlabelled data"/>
<connect from_op="Linear Regression [Sem FS]" from_port="model" to_op="Apply Model Linear Regression [Sem FS]" to_port="model"/>

```
    <connect from_op="Apply Model Linear Regression [Sem FS]" from_port="labelled data"
to_op="Performance Linear Regression [Sem FS]" to_port="labelled data"/>
    <connect from_op="Apply Model Linear Regression [Sem FS]" from_port="model" to_port="result
2"/>
    <connect from_op="Performance Linear Regression [Sem FS]" from_port="performance"
to_port="result 1"/>
    <portSpacing port="source_input 1" spacing="0"/>
    <portSpacing port="sink_result 1" spacing="324"/>
    <portSpacing port="sink_result 2" spacing="0"/>
    <portSpacing port="sink_result 3" spacing="684"/>
  </process>
 </operator>
</process>
```


Modelos M5


```
        <?xml version="1.0" encoding="UTF-8" standalone="no"?>
        <process version="5.0">
          <context>
            <input/>
            <output/>
            <macros/>
          </context>
          <operator activated="true" class="process" compatibility="5.0.11" expanded="true"
name="Process">
            <process expanded="true" height="541" width="1103">
              <operator activated="true" class="retrieve" compatibility="5.0.11" expanded="true"
height="60" name="Retrieve" width="90" x="45" y="30">
                <parameter key="repository_entry" value="BD 20101113"/>
              </operator>
              <operator activated="true" class="remove_duplicates" compatibility="5.0.11"
expanded="true" height="76" name="Remove Duplicates (2)" width="90" x="45" y="120">
                <parameter key="attribute_filter_type" value="subset"/>
                <parameter key="attributes" value="MUNICIPIO|PRECO|MICROZONA RECODE
COD|MACROZONA|V01_AREA|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERV
ACAO_USADO_CONSERVACAO&amp;amp;IDADE|V05_PRESERVACAO_NOVO"/>
              </operator>
              <operator activated="true" class="work_on_subset" compatibility="5.0.11" expanded="true"
height="76" name="Work on Subset" width="90" x="45" y="210">
                <parameter key="attribute_filter_type" value="subset"/>
                <parameter key="attributes"
value="V47_POT_AMENIDADE_UTILIDADES_T3|V46_POT_AMENIDADE_UTILIDADES_T2|V4
5_POT_AMENIDADE_UTILIDADES_T1|V44_POT_AMENIDADE__ZI_T3|V43_POT_AMENIDAD
E__ZI_T2|V42_POT_AMENIDADE_SERVSAUDE_T3|V41_POT_AMENIDADE_SERVSAUDE_T2|
V40_POT_AMENIDADE_SERVSAUDE_T1|V39_POT_AMENIDADE_AREASVERDES_T1|V38_PO
T_AMENIDADE_MOBILIDADE_T2|V37_POT_AMENIDADE_MOBILIDADE_T1|V36_POT_AME
NIDADE_ELEMTURISTICOS_T1|V35_POT_AMENIDADE_EDUCACAO_T3|V34_POT_AMENIDA
DE_EDUCACAO_T2|V33_POT_AMENIDADE_EDUCACAO_T1|V32_POT_AMENIDADE_DIVER
TIMENTOS_T3|V31_POT_AMENIDADE_DIVERTIMENTOS_T2|V30_POT_AMENIDADE_DESPO
RTO_T3|V29_POT_AMENIDADE_DESPORTO_T2|V28_POT_AMENIDADE_DESPORTO_T1|V27_
POT_AMENIDADE_CULTURA_T3|V26_POT_AMENIDADE_CULTURA_T2|V25_POT_AMENIDA
DE_CULTURA_T1|V24_POT_AMENIDADE_COMERCIO_T3|V23_POT_AMENIDADE_COMERCI
O_T2|V22_POT_AMENIDADE_COMERCIO_T1|V21_ATRB_13_HIDROMASSAGEM|V20_ATRB_
12_LAREIRA|V19_ATRB_11_REMODELADO|V18_ATRB_10_WC|V17_ATRB_9_AQUECIMENT
```

O|V16_ATRB_8_PISO|V15_ATRB_7_GARAGEM|V14_ATRB_6_LUGARGARAGEM|V13_ATRB_5_TERRACO|V12_ATRB_4_SOTAO|V11_ATRB_3_VARANDA|V10_ATRB_2_ARREC|V09_ATRIB_1_DUPLEX|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CONSERVACAO&IDADE|V05_PRESERVACAO_NOVO|V04_MACROZONA PRAIAS COD|V03_MACROZONA CENTRO ILH COD|V02_MACROZONA CENTRO AVR COD|V01_AREA|V00_PRECO_M2|MICROZONA RECODE COD|ID"/>
<parameter key="include_special_attributes" value="true"/>
<parameter key="keep_subset_only" value="true"/>
<process expanded="true">
<operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" name="Set Role">
<parameter key="name" value="ID"/>
<parameter key="target_role" value="id"/>
</operator>
<operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" name="Set Role (2)">
<parameter key="name" value="MICROZONA RECODE COD"/>
<parameter key="target_role" value="batch"/>
</operator>
<operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" name="Set Role (3)">
<parameter key="name" value="V00_PRECO_M2"/>
<parameter key="target_role" value="label"/>
</operator>
<connect from_port="exampleSet" to_op="Set Role" to_port="example set input"/>
<connect from_op="Set Role" from_port="example set output" to_op="Set Role (2)" to_port="example set input"/>
<connect from_op="Set Role (2)" from_port="example set output" to_op="Set Role (3)" to_port="example set input"/>
<connect from_op="Set Role (3)" from_port="example set output" to_port="example set"/>
<portSpacing port="source_exampleSet" spacing="0"/>
<portSpacing port="sink_example set" spacing="0"/>
<portSpacing port="sink_through 1" spacing="0"/>
</process>
</operator>
<operator activated="true" class="normalize" compatibility="5.0.11" expanded="true" height="94" name="Normalize" width="90" x="45" y="300">
<parameter key="attribute_filter_type" value="subset"/>
<parameter key="attributes" value="V47_POT_AMENIDADE_UTILIDADES_T3|V46_POT_AMENIDADE_UTILIDADES_T2|V45_POT_AMENIDADE_UTILIDADES_T1|V44_POT_AMENIDADE__ZI_T3|V43_POT_AMENIDADE__ZI_T2|V42_POT_AMENIDADE_SERVSAUDE_T3|V41_POT_AMENIDADE_SERVSAUDE_T2|V40_POT_AMENIDADE_SERVSAUDE_T1|V39_POT_AMENIDADE_AREASVERDES_T1|V38_POT_AMENIDADE_MOBILIDADE_T2|V37_POT_AMENIDADE_MOBILIDADE_T1|V36_POT_AMENIDADE_ELEMTURISTICOS_T1|V35_POT_AMENIDADE_EDUCACAO_T3|V34_POT_AMENIDADE_EDUCACAO_T2|V33_POT_AMENIDADE_EDUCACAO_T1|V32_POT_AMENIDADE_DIVERTIMENTOS_T3|V31_POT_AMENIDADE_DIVERTIMENTOS_T2|V30_POT_AMENIDADE_DESPORTO_T3|V29_POT_AMENIDADE_DESPORTO_T2|V28_POT_AMENIDADE_DESPORTO_T1|V27_POT_AMENIDADE_CULTURA_T3|V26_POT_AMENIDADE_CULTURA_T2|V25_POT_AMENIDADE_CULTURA_T1|V24_POT_AMENIDADE_COMERCIO_T3|V23_POT_AMENIDADE_COMERCIO_T2|V22_POT_AMENIDADE_COMERCIO_T1|V21_ATRB_13_HIDROMASSAGEM|V20_ATRB_12_LAREIRA|V19_ATRB_11_REMODELADO|V18_ATRB_10_WC|V17_ATRB_9_AQUECIMENTO|V16_ATRB_8_PISO|V15_ATRB_7_GARAGEM|V14_ATRB_6_LUGARGARAGEM|V13_ATRB_5_TERRACO|V12_ATRB_4_SOTAO|V11_ATRB_3_VARANDA|V10_ATRB_2_ARREC|V09_ATRIB_1_DUPLEX|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CONSERVACAO&IDADE|V05_PRESERVACAO_NOVO|V04_MACROZONA PRAIAS COD|V03_MACROZONA CENTRO ILH COD|V02_MACROZONA CENTRO AVR COD|V01_AREA"/>
</operator>

```
<operator activated="true" class="remove_correlated_attributes" compatibility="5.0.11" expanded="true" height="76" name="Remove Correlated Attributes" width="90" x="45" y="435"/>
<operator activated="true" class="optimize_selection" compatibility="5.0.11" expanded="true" height="94" name="Optimize Selection" width="90" x="179" y="165">
<parameter key="selection_direction" value="backward"/>
<process expanded="true" height="393" width="567">
<operator activated="true" class="weka:performance_cfs" compatibility="5.0.1" expanded="true" height="76" name="Performance" width="90" x="247" y="30"/>
<connect from_port="example set" to_op="Performance" to_port="example set"/>
<connect from_op="Performance" from_port="performance" to_port="performance"/>
<portSpacing port="source_example set" spacing="0"/>
<portSpacing port="source_through 1" spacing="0"/>
<portSpacing port="sink_performance" spacing="0"/>
</process>
</operator>
<operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true" height="94" name="Select by Weights" width="90" x="380" y="165">
<parameter key="weight" value="0.5"/>
</operator>
<operator activated="true" class="split_data" compatibility="5.0.11" expanded="true" height="94" name="Split Data" width="90" x="514" y="165">
<enumeration key="partitions">
<parameter key="ratio" value="0.7"/>
<parameter key="ratio" value="0.3"/>
</enumeration>
<parameter key="sampling_type" value="stratified sampling"/>
</operator>
<operator activated="true" class="linear_regression" compatibility="5.0.11" expanded="true" height="94" name="Linear Regression [Sem FS]" width="90" x="715" y="165">
<parameter key="feature_selection" value="none"/>
</operator>
<operator activated="true" class="apply_model" compatibility="5.0.11" expanded="true" height="76" name="Apply Model Linear Regression [Sem FS]" width="90" x="849" y="165">
<list key="application_parameters"/>
</operator>
<operator activated="true" class="performance_regression" compatibility="5.0.11" expanded="true" height="76" name="Performance Linear Regression [Sem FS]" width="90" x="983" y="165">
<parameter key="root_mean_squared_error" value="false"/>
<parameter key="squared_correlation" value="true"/>
</operator>
<connect from_op="Retrieve" from_port="output" to_op="Remove Duplicates (2)" to_port="example set input"/>
<connect from_op="Remove Duplicates (2)" from_port="example set output" to_op="Work on Subset" to_port="example set"/>
<connect from_op="Work on Subset" from_port="example set" to_op="Normalize" to_port="example set input"/>
<connect from_op="Normalize" from_port="example set output" to_op="Remove Correlated Attributes" to_port="example set input"/>
<connect from_op="Remove Correlated Attributes" from_port="example set output" to_op="Optimize Selection" to_port="example set in"/>
<connect from_op="Optimize Selection" from_port="example set out" to_op="Select by Weights" to_port="example set input"/>
<connect from_op="Optimize Selection" from_port="weights" to_op="Select by Weights" to_port="weights"/>
<connect from_op="Optimize Selection" from_port="performance" to_port="result 1"/>
<connect from_op="Select by Weights" from_port="example set output" to_op="Split Data" to_port="example set"/>
<connect from_op="Select by Weights" from_port="weights" to_port="result 2"/>
```

```
            <connect from_op="Split Data" from_port="partition 1" to_op="Linear Regression [Sem FS]" to_port="training set"/>
            <connect from_op="Split Data" from_port="partition 2" to_op="Apply Model Linear Regression [Sem FS]" to_port="unlabelled data"/>
            <connect from_op="Linear Regression [Sem FS]" from_port="model" to_op="Apply Model Linear Regression [Sem FS]" to_port="model"/>
            <connect from_op="Linear Regression [Sem FS]" from_port="weights" to_port="result 5"/>
            <connect from_op="Apply Model Linear Regression [Sem FS]" from_port="labelled data" to_op="Performance Linear Regression [Sem FS]" to_port="labelled data"/>
            <connect from_op="Apply Model Linear Regression [Sem FS]" from_port="model" to_port="result 4"/>
            <connect from_op="Performance Linear Regression [Sem FS]" from_port="performance" to_port="result 3"/>
            <portSpacing port="source_input 1" spacing="0"/>
            <portSpacing port="sink_result 1" spacing="234"/>
            <portSpacing port="sink_result 2" spacing="0"/>
            <portSpacing port="sink_result 3" spacing="0"/>
            <portSpacing port="sink_result 4" spacing="0"/>
            <portSpacing port="sink_result 5" spacing="0"/>
            <portSpacing port="sink_result 6" spacing="0"/>
          </process>
        </operator>
</process>
```


Modelos M6


```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.0">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.0.11" expanded="true" name="Process">
    <process expanded="true" height="647" width="1438">
      <operator activated="true" class="retrieve" compatibility="5.0.11" expanded="true" height="60" name="Retrieve" width="90" x="45" y="30">
        <parameter key="repository_entry" value="BD 20101113"/>
      </operator>
      <operator activated="true" class="remove_duplicates" compatibility="5.0.11" expanded="true" height="76" name="Remove Duplicates (2)" width="90" x="45" y="120">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value="MUNICIPIO|PRECO|MICROZONA RECODE COD|MACROZONA|V01_AREA|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CONSERVACAO&amp;amp;IDADE|V05_PRESERVACAO_NOVO"/>
      </operator>
      <operator activated="true" class="work_on_subset" compatibility="5.0.11" expanded="true" height="76" name="Work on Subset" width="90" x="45" y="210">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value="V47_POT_AMENIDADE_UTILIDADES_T3|V46_POT_AMENIDADE_UTILIDADES_T2|V45_POT_AMENIDADE_UTILIDADES_T1|V44_POT_AMENIDADE__ZI_T3|V43_POT_AMENIDADE__ZI_T2|V42_POT_AMENIDADE_SERVSAUDE_T3|V41_POT_AMENIDADE_SERVSAUDE_T2|
```

V40_POT_AMENIDADE_SERVSAUDE_T1|V39_POT_AMENIDADE_AREASVERDES_T1|V38_PO
T_AMENIDADE_MOBILIDADE_T2|V37_POT_AMENIDADE_MOBILIDADE_T1|V36_POT_AME
NIDADE_ELEMTURISTICOS_T1|V35_POT_AMENIDADE_EDUCACAO_T3|V34_POT_AMENIDA
DE_EDUCACAO_T2|V33_POT_AMENIDADE_EDUCACAO_T1|V32_POT_AMENIDADE_DIVER
TIMENTOS_T3|V31_POT_AMENIDADE_DIVERTIMENTOS_T2|V30_POT_AMENIDADE_DESPO
RTO_T3|V29_POT_AMENIDADE_DESPORTO_T2|V28_POT_AMENIDADE_DESPORTO_T1|V27_
POT_AMENIDADE_CULTURA_T3|V26_POT_AMENIDADE_CULTURA_T2|V25_POT_AMENIDA
DE_CULTURA_T1|V24_POT_AMENIDADE_COMERCIO_T3|V23_POT_AMENIDADE_COMERCI
O_T2|V22_POT_AMENIDADE_COMERCIO_T1|V21_ATRB_13_HIDROMASSAGEM|V20_ATRB_
12_LAREIRA|V19_ATRB_11_REMODELADO|V18_ATRB_10_WC|V17_ATRB_9_AQUECIMENT
O|V16_ATRB_8_PISO|V15_ATRB_7_GARAGEM|V14_ATRB_6_LUGARGARAGEM|V13_ATRB_5
_TERRACO|V12_ATRB_4_SOTAO|V11_ATRB_3_VARANDA|V10_ATRB_2_ARREC|V09_ATRIB
_1_DUPLEX|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CO
NSERVACAO&IDADE|V05_PRESERVACAO_NOVO|V04_MACROZONA PRAIAS
COD|V03_MACROZONA CENTRO ILH COD|V02_MACROZONA CENTRO AVR
COD|V01_AREA|V00_PRECO_M2|MICROZONA RECODE COD|ID"/>
<parameter key="include_special_attributes" value="true"/>
<parameter key="keep_subset_only" value="true"/>
<process expanded="true">
<operator activated="true" class="set_role" compatibility="5.0.11" expanded="true"
name="Set Role">
<parameter key="name" value="ID"/>
<parameter key="target_role" value="id"/>
</operator>
<operator activated="true" class="set_role" compatibility="5.0.11" expanded="true"
name="Set Role (2)">
<parameter key="name" value="MICROZONA RECODE COD"/>
<parameter key="target_role" value="batch"/>
</operator>
<operator activated="true" class="set_role" compatibility="5.0.11" expanded="true"
name="Set Role (3)">
<parameter key="name" value="V00_PRECO_M2"/>
<parameter key="target_role" value="label"/>
</operator>
<connect from_port="exampleSet" to_op="Set Role" to_port="example set input"/>
<connect from_op="Set Role" from_port="example set output" to_op="Set Role (2)"
to_port="example set input"/>
<connect from_op="Set Role (2)" from_port="example set output" to_op="Set Role (3)"
to_port="example set input"/>
<connect from_op="Set Role (3)" from_port="example set output" to_port="example set"/>
<portSpacing port="source_exampleSet" spacing="0"/>
<portSpacing port="sink_example set" spacing="0"/>
<portSpacing port="sink_through 1" spacing="0"/>
</process>
</operator>
<operator activated="true" class="normalize" compatibility="5.0.11" expanded="true"
height="94" name="Normalize" width="90" x="45" y="300">
<parameter key="attribute_filter_type" value="subset"/>
<parameter key="attributes"
value="V47_POT_AMENIDADE_UTILIDADES_T3|V46_POT_AMENIDADE_UTILIDADES_T2|V4
5_POT_AMENIDADE_UTILIDADES_T1|V44_POT_AMENIDADE__ZI_T3|V43_POT_AMENIDAD
E__ZI_T2|V42_POT_AMENIDADE_SERVSAUDE_T3|V41_POT_AMENIDADE_SERVSAUDE_T2|
V40_POT_AMENIDADE_SERVSAUDE_T1|V39_POT_AMENIDADE_AREASVERDES_T1|V38_PO
T_AMENIDADE_MOBILIDADE_T2|V37_POT_AMENIDADE_MOBILIDADE_T1|V36_POT_AME
NIDADE_ELEMTURISTICOS_T1|V35_POT_AMENIDADE_EDUCACAO_T3|V34_POT_AMENIDA
DE_EDUCACAO_T2|V33_POT_AMENIDADE_EDUCACAO_T1|V32_POT_AMENIDADE_DIVER
TIMENTOS_T3|V31_POT_AMENIDADE_DIVERTIMENTOS_T2|V30_POT_AMENIDADE_DESPO
RTO_T3|V29_POT_AMENIDADE_DESPORTO_T2|V28_POT_AMENIDADE_DESPORTO_T1|V27_
POT_AMENIDADE_CULTURA_T3|V26_POT_AMENIDADE_CULTURA_T2|V25_POT_AMENIDA
DE_CULTURA_T1|V24_POT_AMENIDADE_COMERCIO_T3|V23_POT_AMENIDADE_COMERCI

O_T2|V22_POT_AMENIDADE_COMERCIO_T1|V21_ATRB_13_HIDROMASSAGEM|V20_ATRB_12_LAREIRA|V19_ATRB_11_REMODELADO|V18_ATRB_10_WC|V17_ATRB_9_AQUECIMENTO|V16_ATRB_8_PISO|V15_ATRB_7_GARAGEM|V14_ATRB_6_LUGARGARAGEM|V13_ATRB_5_TERRACO|V12_ATRB_4_SOTAO|V11_ATRB_3_VARANDA|V10_ATRB_2_ARREC|V09_ATRIB_1_DUPLEX|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CONSERVACAO&IDADE|V05_PRESERVACAO_NOVO|V04_MACROZONA PRAIAS COD|V03_MACROZONA CENTRO ILH COD|V02_MACROZONA CENTRO AVR COD|V01_AREA"/>
</operator>
<operator activated="true" class="remove_correlated_attributes" compatibility="5.0.11" expanded="true" height="76" name="Remove Correlated Attributes" width="90" x="45" y="435"/>
<operator activated="true" class="optimize_selection" compatibility="5.0.11" expanded="true" height="94" name="Optimize Selection" width="90" x="313" y="165">
<process expanded="true" height="393" width="567">
<operator activated="true" class="weka:performance_cfs" compatibility="5.0.1" expanded="true" height="76" name="Performance" width="90" x="247" y="30"/>
<connect from_port="example set" to_op="Performance" to_port="example set"/>
<connect from_op="Performance" from_port="performance" to_port="performance"/>
<portSpacing port="source_example set" spacing="0"/>
<portSpacing port="source_through 1" spacing="0"/>
<portSpacing port="sink_performance" spacing="0"/>
</process>
</operator>
<operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true" height="94" name="Select by Weights" width="90" x="447" y="165">
<parameter key="weight" value="0.5"/>
</operator>
<operator activated="true" class="split_data" compatibility="5.0.11" expanded="true" height="94" name="Split Data" width="90" x="648" y="165">
<enumeration key="partitions">
<parameter key="ratio" value="0.7"/>
<parameter key="ratio" value="0.3"/>
</enumeration>
<parameter key="sampling_type" value="stratified sampling"/>
</operator>
<operator activated="true" class="linear_regression" compatibility="5.0.11" expanded="true" height="94" name="Linear Regression [Sem FS]" width="90" x="849" y="165">
<parameter key="feature_selection" value="none"/>
</operator>
<operator activated="true" class="apply_model" compatibility="5.0.11" expanded="true" height="76" name="Apply Model Linear Regression [Sem FS]" width="90" x="983" y="165">
<list key="application_parameters"/>
</operator>
<operator activated="true" class="performance_regression" compatibility="5.0.11" expanded="true" height="76" name="Performance Linear Regression [Sem FS]" width="90" x="1117" y="165">
<parameter key="root_mean_squared_error" value="false"/>
<parameter key="squared_correlation" value="true"/>
</operator>
<connect from_op="Retrieve" from_port="output" to_op="Remove Duplicates (2)" to_port="example set input"/>
<connect from_op="Remove Duplicates (2)" from_port="example set output" to_op="Work on Subset" to_port="example set"/>
<connect from_op="Work on Subset" from_port="example set" to_op="Normalize" to_port="example set input"/>
<connect from_op="Normalize" from_port="example set output" to_op="Remove Correlated Attributes" to_port="example set input"/>
<connect from_op="Remove Correlated Attributes" from_port="example set output" to_op="Optimize Selection" to_port="example set in"/>

```
            <connect from_op="Optimize Selection" from_port="example set out" to_op="Select by Weights" to_port="example set input"/>
            <connect from_op="Optimize Selection" from_port="weights" to_op="Select by Weights" to_port="weights"/>
            <connect from_op="Optimize Selection" from_port="performance" to_port="result 1"/>
            <connect from_op="Select by Weights" from_port="example set output" to_op="Split Data" to_port="example set"/>
            <connect from_op="Select by Weights" from_port="weights" to_port="result 2"/>
            <connect from_op="Split Data" from_port="partition 1" to_op="Linear Regression [Sem FS]" to_port="training set"/>
            <connect from_op="Split Data" from_port="partition 2" to_op="Apply Model Linear Regression [Sem FS]" to_port="unlabelled data"/>
            <connect from_op="Linear Regression [Sem FS]" from_port="model" to_op="Apply Model Linear Regression [Sem FS]" to_port="model"/>
            <connect from_op="Linear Regression [Sem FS]" from_port="weights" to_port="result 5"/>
            <connect from_op="Apply Model Linear Regression [Sem FS]" from_port="labelled data" to_op="Performance Linear Regression [Sem FS]" to_port="labelled data"/>
            <connect from_op="Apply Model Linear Regression [Sem FS]" from_port="model" to_port="result 4"/>
            <connect from_op="Performance Linear Regression [Sem FS]" from_port="performance" to_port="result 3"/>
            <portSpacing port="source_input 1" spacing="0"/>
            <portSpacing port="sink_result 1" spacing="180"/>
            <portSpacing port="sink_result 2" spacing="0"/>
            <portSpacing port="sink_result 3" spacing="0"/>
            <portSpacing port="sink_result 4" spacing="0"/>
            <portSpacing port="sink_result 5" spacing="0"/>
            <portSpacing port="sink_result 6" spacing="0"/>
          </process>
        </operator>
</process>
```

Modelos M7


```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.0">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.0.11" expanded="true" name="Process">
    <process expanded="true" height="514" width="1170">
      <operator activated="true" class="retrieve" compatibility="5.0.11" expanded="true" height="60" name="Retrieve" width="90" x="45" y="30">
        <parameter key="repository_entry" value="BD 20101113"/>
      </operator>
      <operator activated="true" class="remove_duplicates" compatibility="5.0.11" expanded="true" height="76" name="Remove Duplicates (2)" width="90" x="45" y="120">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value="MUNICIPIO|PRECO|MICROZONA RECODE COD|MACROZONA|V01_AREA|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CONSERVACAO&amp;amp;IDADE|V05_PRESERVACAO_NOVO"/>
      </operator>
      <operator activated="true" class="work_on_subset" compatibility="5.0.11" expanded="true" height="76" name="Work on Subset" width="90" x="45" y="210">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value="V47_POT_AMENIDADE_UTILIDADES_T3|V46_POT_AMENIDADE_UTILIDADES_T2|V45_POT_AMENIDADE_UTILIDADES_T1|V44_POT_AMENIDADE__ZI_T3|V43_POT_AMENIDADE__ZI_T2|V42_POT_AMENIDADE_SERVSAUDE_T3|V41_POT_AMENIDADE_SERVSAUDE_T2|V40_POT_AMENIDADE_SERVSAUDE_T1|V39_POT_AMENIDADE_AREASVERDES_T1|V38_POT_AMENIDADE_MOBILIDADE_T2|V37_POT_AMENIDADE_MOBILIDADE_T1|V36_POT_AMENIDADE_ELEMTURISTICOS_T1|V35_POT_AMENIDADE_EDUCACAO_T3|V34_POT_AMENIDADE_EDUCACAO_T2|V33_POT_AMENIDADE_EDUCACAO_T1|V32_POT_AMENIDADE_DIVERTIMENTOS_T3|V31_POT_AMENIDADE_DIVERTIMENTOS_T2|V30_POT_AMENIDADE_DESPORTO_T3|V29_POT_AMENIDADE_DESPORTO_T2|V28_POT_AMENIDADE_DESPORTO_T1|V27_POT_AMENIDADE_CULTURA_T3|V26_POT_AMENIDADE_CULTURA_T2|V25_POT_AMENIDADE_CULTURA_T1|V24_POT_AMENIDADE_COMERCIO_T3|V23_POT_AMENIDADE_COMERCIO_T2|V22_POT_AMENIDADE_COMERCIO_T1|V21_ATRB_13_HIDROMASSAGEM|V20_ATRB_12_LAREIRA|V19_ATRB_11_REMODELADO|V18_ATRB_10_WC|V17_ATRB_9_AQUECIMENTO|V16_ATRB_8_PISO|V15_ATRB_7_GARAGEM|V14_ATRB_6_LUGARGARAGEM|V13_ATRB_5_TERRACO|V12_ATRB_4_SOTAO|V11_ATRB_3_VARANDA|V10_ATRB_2_ARREC|V09_ATRIB_1_DUPLEX|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CONSERVACAO&amp;amp;IDADE|V05_PRESERVACAO_NOVO|V04_MACROZONA PRAIAS COD|V03_MACROZONA CENTRO ILH COD|V02_MACROZONA CENTRO AVR COD|V01_AREA|V00_PRECO_M2|MICROZONA RECODE COD|ID"/>
        <parameter key="include_special_attributes" value="true"/>
        <parameter key="keep_subset_only" value="true"/>
        <process expanded="true">
          <operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" name="Set Role">
            <parameter key="name" value="ID"/>
            <parameter key="target_role" value="id"/>
          </operator>
          <operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" name="Set Role (2)">
            <parameter key="name" value="MICROZONA RECODE COD"/>
            <parameter key="target_role" value="batch"/>
          </operator>
```

```
<operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" name="Set Role (3)">
<parameter key="name" value="V00_PRECO_M2"/>
<parameter key="target_role" value="label"/>
</operator>
<connect from_port="exampleSet" to_op="Set Role" to_port="example set input"/>
<connect from_op="Set Role" from_port="example set output" to_op="Set Role (2)" to_port="example set input"/>
<connect from_op="Set Role (2)" from_port="example set output" to_op="Set Role (3)" to_port="example set input"/>
<connect from_op="Set Role (3)" from_port="example set output" to_port="example set"/>
<portSpacing port="source_exampleSet" spacing="0"/>
<portSpacing port="sink_example set" spacing="0"/>
<portSpacing port="sink_through 1" spacing="0"/>
</process>
</operator>
<operator activated="true" class="normalize" compatibility="5.0.11" expanded="true" height="94" name="Normalize (2)" width="90" x="45" y="300">
<parameter key="attribute_filter_type" value="subset"/>
<parameter key="attributes" value="V47_POT_AMENIDADE_UTILIDADES_T3|V46_POT_AMENIDADE_UTILIDADES_T2|V45_POT_AMENIDADE_UTILIDADES_T1|V44_POT_AMENIDADE__ZI_T3|V43_POT_AMENIDADE__ZI_T2|V42_POT_AMENIDADE_SERVSAUDE_T3|V41_POT_AMENIDADE_SERVSAUDE_T2|V40_POT_AMENIDADE_SERVSAUDE_T1|V39_POT_AMENIDADE_AREASVERDES_T1|V38_POT_AMENIDADE_MOBILIDADE_T2|V37_POT_AMENIDADE_MOBILIDADE_T1|V36_POT_AMENIDADE_ELEMTURISTICOS_T1|V35_POT_AMENIDADE_EDUCACAO_T3|V34_POT_AMENIDADE_EDUCACAO_T2|V33_POT_AMENIDADE_EDUCACAO_T1|V32_POT_AMENIDADE_DIVERTIMENTOS_T3|V31_POT_AMENIDADE_DIVERTIMENTOS_T2|V30_POT_AMENIDADE_DESPORTO_T3|V29_POT_AMENIDADE_DESPORTO_T2|V28_POT_AMENIDADE_DESPORTO_T1|V27_POT_AMENIDADE_CULTURA_T3|V26_POT_AMENIDADE_CULTURA_T2|V25_POT_AMENIDADE_CULTURA_T1|V24_POT_AMENIDADE_COMERCIO_T3|V23_POT_AMENIDADE_COMERCIO_T2|V22_POT_AMENIDADE_COMERCIO_T1|V21_ATRB_13_HIDROMASSAGEM|V20_ATRB_12_LAREIRA|V19_ATRB_11_REMODELADO|V18_ATRB_10_WC|V17_ATRB_9_AQUECIMENTO|V16_ATRB_8_PISO|V15_ATRB_7_GARAGEM|V14_ATRB_6_LUGARGARAGEM|V13_ATRB_5_TERRACO|V12_ATRB_4_SOTAO|V11_ATRB_3_VARANDA|V10_ATRB_2_ARREC|V09_ATRIB_1_DUPLEX|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CONSERVACAO&amp;amp;IDADE|V05_PRESERVACAO_NOVO|V04_MACROZONA PRAIAS COD|V03_MACROZONA CENTRO ILH COD|V02_MACROZONA CENTRO AVR COD|V01_AREA"/>
</operator>
<operator activated="true" class="weight_by_pca" compatibility="5.0.11" expanded="true" height="76" name="Weight by PCA" width="90" x="246" y="165"/>
<operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true" height="94" name="Select by Weights" width="90" x="380" y="165">
<parameter key="weight" value="0.5"/>
</operator>
<operator activated="true" class="split_data" compatibility="5.0.11" expanded="true" height="94" name="Split Data" width="90" x="581" y="165">
<enumeration key="partitions">
<parameter key="ratio" value="0.7"/>
<parameter key="ratio" value="0.3"/>
</enumeration>
<parameter key="sampling_type" value="stratified sampling"/>
</operator>
<operator activated="true" class="linear_regression" compatibility="5.0.11" expanded="true" height="94" name="Linear Regression [Sem FS]" width="90" x="782" y="165">
<parameter key="feature_selection" value="none"/>
</operator>
<operator activated="true" class="apply_model" compatibility="5.0.11" expanded="true" height="76" name="Apply Model Linear Regression [Sem FS]" width="90" x="916" y="165">
```

```
            <list key="application_parameters"/>
            <parameter key="create_view" value="true"/>
          </operator>
          <operator activated="true" class="performance_regression" compatibility="5.0.11" expanded="true" height="76" name="Performance Linear Regression [Sem FS]" width="90" x="1050" y="165">
            <parameter key="root_mean_squared_error" value="false"/>
            <parameter key="squared_correlation" value="true"/>
          </operator>
          <connect from_op="Retrieve" from_port="output" to_op="Remove Duplicates (2)" to_port="example set input"/>
          <connect from_op="Remove Duplicates (2)" from_port="example set output" to_op="Work on Subset" to_port="example set"/>
          <connect from_op="Work on Subset" from_port="example set" to_op="Normalize (2)" to_port="example set input"/>
          <connect from_op="Normalize (2)" from_port="example set output" to_op="Weight by PCA" to_port="example set"/>
          <connect from_op="Weight by PCA" from_port="weights" to_op="Select by Weights" to_port="weights"/>
          <connect from_op="Weight by PCA" from_port="example set" to_op="Select by Weights" to_port="example set input"/>
          <connect from_op="Select by Weights" from_port="example set output" to_op="Split Data" to_port="example set"/>
          <connect from_op="Select by Weights" from_port="weights" to_port="result 1"/>
          <connect from_op="Split Data" from_port="partition 1" to_op="Linear Regression [Sem FS]" to_port="training set"/>
          <connect from_op="Split Data" from_port="partition 2" to_op="Apply Model Linear Regression [Sem FS]" to_port="unlabelled data"/>
          <connect from_op="Linear Regression [Sem FS]" from_port="model" to_op="Apply Model Linear Regression [Sem FS]" to_port="model"/>
          <connect from_op="Linear Regression [Sem FS]" from_port="weights" to_port="result 4"/>
          <connect from_op="Apply Model Linear Regression [Sem FS]" from_port="labelled data" to_op="Performance Linear Regression [Sem FS]" to_port="labelled data"/>
          <connect from_op="Apply Model Linear Regression [Sem FS]" from_port="model" to_port="result 3"/>
          <connect from_op="Performance Linear Regression [Sem FS]" from_port="performance" to_port="result 2"/>
          <portSpacing port="source_input 1" spacing="0"/>
          <portSpacing port="sink_result 1" spacing="90"/>
          <portSpacing port="sink_result 2" spacing="36"/>
          <portSpacing port="sink_result 3" spacing="0"/>
          <portSpacing port="sink_result 4" spacing="18"/>
          <portSpacing port="sink_result 5" spacing="0"/>
        </process>
      </operator>
</process>
```

Modelos M8


```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.0">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.0.11" expanded="true" name="Process">
    <process expanded="true" height="521" width="1237">
      <operator activated="true" class="retrieve" compatibility="5.0.11" expanded="true" height="60" name="Retrieve" width="90" x="45" y="30">
        <parameter key="repository_entry" value="BD 20101113"/>
      </operator>
      <operator activated="true" class="remove_duplicates" compatibility="5.0.11" expanded="true" height="76" name="Remove Duplicates (2)" width="90" x="45" y="120">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value="MUNICIPIO|PRECO|MICROZONA RECODE COD|MACROZONA|V01_AREA|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CONSERVACAO&amp;amp;IDADE|V05_PRESERVACAO_NOVO"/>
      </operator>
      <operator activated="true" class="work_on_subset" compatibility="5.0.11" expanded="true" height="76" name="Work on Subset" width="90" x="45" y="210">
        <parameter key="attribute_filter_type" value="subset"/>
        <parameter key="attributes" value="V47_POT_AMENIDADE_UTILIDADES_T3|V46_POT_AMENIDADE_UTILIDADES_T2|V45_POT_AMENIDADE_UTILIDADES_T1|V44_POT_AMENIDADE__ZI_T3|V43_POT_AMENIDADE__ZI_T2|V42_POT_AMENIDADE_SERVSAUDE_T3|V41_POT_AMENIDADE_SERVSAUDE_T2|V40_POT_AMENIDADE_SERVSAUDE_T1|V39_POT_AMENIDADE_AREASVERDES_T1|V38_POT_AMENIDADE_MOBILIDADE_T2|V37_POT_AMENIDADE_MOBILIDADE_T1|V36_POT_AMENIDADE_ELEMTURISTICOS_T1|V35_POT_AMENIDADE_EDUCACAO_T3|V34_POT_AMENIDADE_EDUCACAO_T2|V33_POT_AMENIDADE_EDUCACAO_T1|V32_POT_AMENIDADE_DIVERTIMENTOS_T3|V31_POT_AMENIDADE_DIVERTIMENTOS_T2|V30_POT_AMENIDADE_DESPORTO_T3|V29_POT_AMENIDADE_DESPORTO_T2|V28_POT_AMENIDADE_DESPORTO_T1|V27_POT_AMENIDADE_CULTURA_T3|V26_POT_AMENIDADE_CULTURA_T2|V25_POT_AMENIDADE_CULTURA_T1|V24_POT_AMENIDADE_COMERCIO_T3|V23_POT_AMENIDADE_COMERCIO_T2|V22_POT_AMENIDADE_COMERCIO_T1|V21_ATRB_13_HIDROMASSAGEM|V20_ATRB_12_LAREIRA|V19_ATRB_11_REMODELADO|V18_ATRB_10_WC|V17_ATRB_9_AQUECIMENTO|V16_ATRB_8_PISO|V15_ATRB_7_GARAGEM|V14_ATRB_6_LUGARGARAGEM|V13_ATRB_5_TERRACO|V12_ATRB_4_SOTAO|V11_ATRB_3_VARANDA|V10_ATRB_2_ARREC|V09_ATRIB_1_DUPLEX|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CONSERVACAO&amp;amp;IDADE|V05_PRESERVACAO_NOVO|V04_MACROZONA PRAIAS COD|V03_MACROZONA CENTRO ILH COD|V02_MACROZONA CENTRO AVR COD|V01_AREA|V00_PRECO_M2|MICROZONA RECODE COD|ID"/>
        <parameter key="include_special_attributes" value="true"/>
        <parameter key="keep_subset_only" value="true"/>
        <process expanded="true">
          <operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" name="Set Role">
            <parameter key="name" value="ID"/>
            <parameter key="target_role" value="id"/>
          </operator>
          <operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" name="Set Role (2)">
            <parameter key="name" value="MICROZONA RECODE COD"/>
            <parameter key="target_role" value="batch"/>
          </operator>
```

```
            <operator activated="true" class="set_role" compatibility="5.0.11" expanded="true" name="Set Role (3)">
              <parameter key="name" value="V00_PRECO_M2"/>
              <parameter key="target_role" value="label"/>
            </operator>
            <connect from_port="exampleSet" to_op="Set Role" to_port="example set input"/>
            <connect from_op="Set Role" from_port="example set output" to_op="Set Role (2)" to_port="example set input"/>
            <connect from_op="Set Role (2)" from_port="example set output" to_op="Set Role (3)" to_port="example set input"/>
            <connect from_op="Set Role (3)" from_port="example set output" to_port="example set"/>
            <portSpacing port="source_exampleSet" spacing="0"/>
            <portSpacing port="sink_example set" spacing="0"/>
            <portSpacing port="sink_through 1" spacing="0"/>
          </process>
        </operator>
        <operator activated="true" class="normalize" compatibility="5.0.11" expanded="true" height="94" name="Normalize (2)" width="90" x="45" y="300">
          <parameter key="attribute_filter_type" value="subset"/>
          <parameter key="attributes" value="V47_POT_AMENIDADE_UTILIDADES_T3|V46_POT_AMENIDADE_UTILIDADES_T2|V45_POT_AMENIDADE_UTILIDADES_T1|V44_POT_AMENIDADE__ZI_T3|V43_POT_AMENIDADE__ZI_T2|V42_POT_AMENIDADE_SERVSAUDE_T3|V41_POT_AMENIDADE_SERVSAUDE_T2|V40_POT_AMENIDADE_SERVSAUDE_T1|V39_POT_AMENIDADE_AREASVERDES_T1|V38_POT_AMENIDADE_MOBILIDADE_T2|V37_POT_AMENIDADE_MOBILIDADE_T1|V36_POT_AMENIDADE_ELEMTURISTICOS_T1|V35_POT_AMENIDADE_EDUCACAO_T3|V34_POT_AMENIDADE_EDUCACAO_T2|V33_POT_AMENIDADE_EDUCACAO_T1|V32_POT_AMENIDADE_DIVERTIMENTOS_T3|V31_POT_AMENIDADE_DIVERTIMENTOS_T2|V30_POT_AMENIDADE_DESPORTO_T3|V29_POT_AMENIDADE_DESPORTO_T2|V28_POT_AMENIDADE_DESPORTO_T1|V27_POT_AMENIDADE_CULTURA_T3|V26_POT_AMENIDADE_CULTURA_T2|V25_POT_AMENIDADE_CULTURA_T1|V24_POT_AMENIDADE_COMERCIO_T3|V23_POT_AMENIDADE_COMERCIO_T2|V22_POT_AMENIDADE_COMERCIO_T1|V21_ATRB_13_HIDROMASSAGEM|V20_ATRB_12_LAREIRA|V19_ATRB_11_REMODELADO|V18_ATRB_10_WC|V17_ATRB_9_AQUECIMENTO|V16_ATRB_8_PISO|V15_ATRB_7_GARAGEM|V14_ATRB_6_LUGARGARAGEM|V13_ATRB_5_TERRACO|V12_ATRB_4_SOTAO|V11_ATRB_3_VARANDA|V10_ATRB_2_ARREC|V09_ATRIB_1_DUPLEX|V08_TIPO_MORADIA|V07_TIPOLOGIA_APART|V06_PRESERVACAO_USADO_CONSERVACAO&amp;amp;IDADE|V05_PRESERVACAO_NOVO|V04_MACROZONA PRAIAS COD|V03_MACROZONA CENTRO ILH COD|V02_MACROZONA CENTRO AVR COD|V01_AREA"/>
        </operator>
        <operator activated="true" class="remove_correlated_attributes" compatibility="5.0.11" expanded="true" height="76" name="Remove Correlated Attributes" width="90" x="45" y="435"/>
        <operator activated="true" class="weight_by_svm" compatibility="5.0.11" expanded="true" height="76" name="Weight by SVM" width="90" x="246" y="210"/>
        <operator activated="true" class="select_by_weights" compatibility="5.0.11" expanded="true" height="94" name="Select by Weights" width="90" x="380" y="210">
          <parameter key="weight" value="0.5"/>
        </operator>
        <operator activated="true" class="split_data" compatibility="5.0.11" expanded="true" height="94" name="Split Data" width="90" x="581" y="210">
          <enumeration key="partitions">
            <parameter key="ratio" value="0.7"/>
            <parameter key="ratio" value="0.3"/>
          </enumeration>
          <parameter key="sampling_type" value="stratified sampling"/>
        </operator>
        <operator activated="true" class="linear_regression" compatibility="5.0.11" expanded="true" height="94" name="Linear Regression [Sem FS]" width="90" x="849" y="210">
          <parameter key="feature_selection" value="none"/>
        </operator>
```

<operator activated="true" class="apply_model" compatibility="5.0.11" expanded="true" height="76" name="Apply Model Linear Regression [Sem FS]" width="90" x="983" y="210">
<list key="application_parameters"/>
<parameter key="create_view" value="true"/>
</operator>
<operator activated="true" class="performance_regression" compatibility="5.0.11" expanded="true" height="76" name="Performance Linear Regression [Sem FS]" width="90" x="1117" y="210">
<parameter key="root_mean_squared_error" value="false"/>
<parameter key="squared_correlation" value="true"/>
</operator>
<connect from_op="Retrieve" from_port="output" to_op="Remove Duplicates (2)" to_port="example set input"/>
<connect from_op="Remove Duplicates (2)" from_port="example set output" to_op="Work on Subset" to_port="example set"/>
<connect from_op="Work on Subset" from_port="example set" to_op="Normalize (2)" to_port="example set input"/>
<connect from_op="Normalize (2)" from_port="example set output" to_op="Remove Correlated Attributes" to_port="example set input"/>
<connect from_op="Remove Correlated Attributes" from_port="example set output" to_op="Weight by SVM" to_port="example set"/>
<connect from_op="Weight by SVM" from_port="weights" to_op="Select by Weights" to_port="weights"/>
<connect from_op="Weight by SVM" from_port="example set" to_op="Select by Weights" to_port="example set input"/>
<connect from_op="Select by Weights" from_port="example set output" to_op="Split Data" to_port="example set"/>
<connect from_op="Select by Weights" from_port="weights" to_port="result 1"/>
<connect from_op="Split Data" from_port="partition 1" to_op="Linear Regression [Sem FS]" to_port="training set"/>
<connect from_op="Split Data" from_port="partition 2" to_op="Apply Model Linear Regression [Sem FS]" to_port="unlabelled data"/>
<connect from_op="Linear Regression [Sem FS]" from_port="model" to_op="Apply Model Linear Regression [Sem FS]" to_port="model"/>
<connect from_op="Linear Regression [Sem FS]" from_port="weights" to_port="result 4"/>
<connect from_op="Apply Model Linear Regression [Sem FS]" from_port="labelled data" to_op="Performance Linear Regression [Sem FS]" to_port="labelled data"/>
<connect from_op="Apply Model Linear Regression [Sem FS]" from_port="model" to_port="result 3"/>
<connect from_op="Performance Linear Regression [Sem FS]" from_port="performance" to_port="result 2"/>
<portSpacing port="source_input 1" spacing="0"/>
<portSpacing port="sink_result 1" spacing="126"/>
<portSpacing port="sink_result 2" spacing="36"/>
<portSpacing port="sink_result 3" spacing="0"/>
<portSpacing port="sink_result 4" spacing="0"/>
<portSpacing port="sink_result 5" spacing="0"/>
</process>
</operator>
</process>